

The international journal of science / 12 January 2023

nature

CARRIED AWAY

How grain shape affects fluid transport of sediment

Inequality in India
Marginalized castes can face major barriers in science

Balanced diet
Traditional knowledge offers overlooked route to food security

Uneven inheritance
Damaged paternal DNA skips a generation in nematode worms

VISIT…

LANZAROTE Caliente.com

Nature.2023.01.14

[Sat, 14 Jan 2023]

- This Week
- News in Focus
- Opinion
- Work
- Research

This Week

- **Vaccine incentives do not backfire — policymakers take note** [11 Jan 2023]
  • Rigorously designed studies can dispel suspicions that offering cash for vaccines erodes trust, and help health-care providers to design effective immunization campaigns.
- **Don’t wait for COP: the end of the fossil-fuel age must start now** [11 Jan 2023]
  • UN climate conferences are too beholden to oil and gas interests. Like-minded nations must come together to keep climate hopes alive.
- **Why NASA and federal agencies are declaring this the Year of Open Science** [11 Jan 2023]
  • Here’s how NASA is incentivizing open science, and how you can too.

- EDITORIAL
- 11 January 2023

Vaccine incentives do not backfire — policymakers take note

Rigorously designed studies can dispel suspicions that offering cash for vaccines erodes trust, and help health-care providers to design effective immunization campaigns.

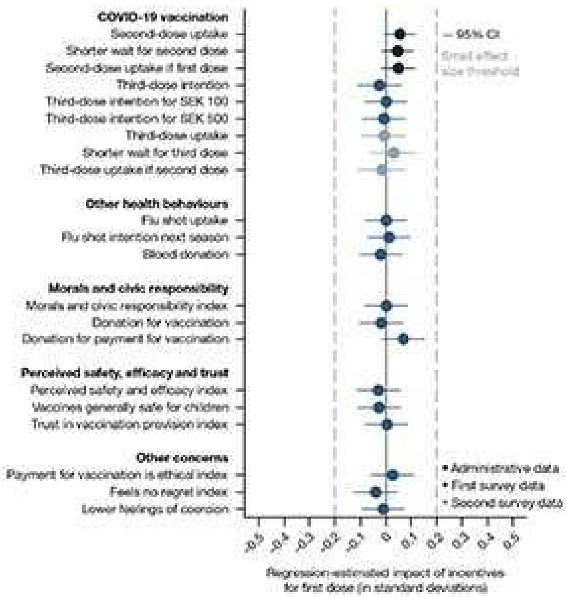


A trial in Sweden showed that uptake rates rose slightly if people were given 200 krona (US$24) to have a COVID-19 vaccine.Credit: Karol Serewis/SOPA Images/LightRocket/Getty

Around the middle of 2021, some governments around the world began to get creative with their COVID-19 vaccination programmes. Eager to get

vaccines into as many arms as possible, they began to offer unusual incentives: in San Luis, Philippines, vaccinated residents could enter a sweepstake to win a cow; in Hong Kong, the windfall was an apartment; in West Virginia, it was a hunting rifle, among other prizes.

Evidence about the effectiveness of such efforts is mixed[1], and the research that has been done suggests that guaranteed, immediate payments are more effective than gimmicks and lotteries[2]. A randomized trial in Sweden in 2021, for example, showed that a cash payment of 200 krona (US$24) to those receiving their first dose of a COVID-19 vaccine was enough to drive up vaccination rates by 4.2 percentage points[3].



Read the paper: Financial incentives for vaccination do not have negative unintended consequences

However, people and policymakers still worry that such incentives could have unintended negative consequences. People might expect payment for future vaccinations, for instance. Or they might think the payments are compensation for something that isn't safe or could be unpleasant. Now, the results of two trials published in a paper on 11 January in *Nature* suggest that simple cash payments have no such unintended effects[4]. The research team took the important step of designing the studies to test a null hypothesis. This gave them the statistical power to detect the absence of an effect, and not just document a lack of evidence for negative consequences.

Such studies (including testing a null result) add to the discussion around financial incentives for vaccination and highlight the value of strong research for policy-relevant questions.

Financial incentives — and disincentives — have long been used to achieve public-health goals, but as a general rule, they have targeted mostly unhealthy behaviours. Take, for example, taxes on cigarettes to discourage smoking. Randomized trials indicate that direct cash payments can be effective in stopping smoking during pregnancy[5].

But there's been a long-standing worry that offering payments for vaccination could reduce people's trust in the safety of vaccines, erode their sense of altruism and make it difficult to implement future vaccine programmes without payments. Although these concerns are largely theoretical, they have led some policy advisers to recommend against implementing programmes that offer cash to promote healthy behaviours. The authors of the *Nature* paper, a team of economists from Europe and the United States, address such concerns head on by analysing data from randomized studies of more than 2,700 people involved in the Swedish cash-incentive programme and of 3,000 people in the United States.



How COVID vaccines shaped 2021 in eight powerful charts

As well as gathering data on second-dose uptake, the Swedish study also surveyed attitudes towards vaccines and other health behaviours. The US study looked at whether informing participants about the existence of incentive programmes that offer public funds for COVID-19 vaccines

affected their perception of the products' safety or efficacy, their sense of civic responsibility or their intention to get an influenza shot or donate blood. Overall, there was no difference in any of these parameters. Nor did participants report feeling coerced by the incentives.

No single experiment can completely answer questions about unintended consequences, and this study, like any other, has its caveats. Ethical requirements dictated that all participants knew they were part of an experiment, and it is possible that this knowledge coloured their responses to survey questions. The results also might not apply to other countries or other vaccines: the COVID-19 pandemic is unique in its pre-eminence as well as its politicization, and both regions studied are high-income, Western countries. Nor does this study address wider and more fundamental ethical questions around paying people to get vaccines.

And although the results support the idea that monetary incentives can be deployed without unintended negative consequences, they do not suggest that such approaches should be prioritized over improving access to vaccines when there are insufficient resources to do both. Efforts to build trust or make it easier for people to receive vaccines — by translating information about inoculations into different languages, for example, or bringing vaccination centres closer to the populations most in need — are well-established ways to boost vaccination rates and, in some cases, could cost less than offering financial rewards.

But for communities that have the resources to both bolster access and provide incentives, the latest work brings important data and a rigorous approach to a discussion that has been heavily influenced by untested assumptions about hypothetical negative consequences. More such studies are needed on effective approaches for policy-relevant questions. In the case of vaccines, such information can be harnessed to help health-care providers design alternative ways to promote vaccination — before another pandemic again entices governments to get ahead of the science and become creative.

Nature **613**, 215 (2023)

doi: https://doi.org/10.1038/d41586-023-00018-z

References

1. Thirumurthy, H., Milkman, K. L., Volpp, K. G., Buttenheim, A. M. & Pope, D. G. *et al. PLoS ONE* **17**, e0263425 (2022).

2. Brewer, N. T. *et al. Lancet Reg. Health Am.* **8**, 100205 (2022).

3. Campos-Mercade, P. *et al. Science* **374**, 879–882 (2021).

4. Schneider, F. H. *et al. Nature* https://doi.org/10.1038/s41586-022-05512-4 (2023).

5. Berlin, I. *et al. BMJ* **375**, e065217 (2021).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00018-z

- EDITORIAL
- 11 January 2023

Don’t wait for COP: the end of the fossil-fuel age must start now

UN climate conferences are too beholden to oil and gas interests. Like-minded nations must come together to keep climate hopes alive.



A cross-party group of UK parliamentarians wants to end new licences for oil and gas drilling in the country and its waters.Credit: Matthew Lloyd/Bloomberg/Getty

The COP27 United Nations climate conference held in Egypt in November was a mixed bag (*Nature* **612**, 16–17; 2022). Although countries recommitted to the goal of the 2015 Paris climate agreement — to limit

global warming to 1.5 °C above pre-industrial temperatures — they made no commitments to phase out fossil fuels. Some even pushed to abandon the 1.5 °C target, saying it is not realistic on the basis of current trends. Thankfully, they were outnumbered.

This could well prove to be a temporary victory, however: the question of burying the 1.5 °C target will re-emerge later this year, at COP28 in Dubai, which is part of a major oil- and gas-producing region. The fight to decarbonize energy supplies is one the world cannot afford to lose.

At present, humanity is on a path to between 2.2 °C and 3.4 °C of warming, with 2.4 °C the most likely value if countries deliver on the emissions reductions they promised to make by 2030, according to the Climate Action Tracker, an independent group of researchers that monitors government action on climate. But temperature increases are likely to be at the higher end of the range if pledges are not kept and business continues as usual.



Why fossil fuel subsidies are so hard to kill

If 1.5 °C is to be achieved, the world must immediately halt new oil and gas development and transition rapidly to renewable energy. That's one of the conclusions of a cross-party report from members of the UK House of Commons, which was published last week (see go.nature.com/3qqkthx). The report draws extensively on evidence from the research community,

including the parliamentarians' expert adviser Jim Watson, the director of the UCL Institute for Sustainable Resources in London.

The need to stop the issuing of new licences for oil and gas development to achieve 1.5 °C is also accepted by the International Energy Agency (IEA), the intergovernmental body established to ensure global energy security. Yet just a handful of nations and regions have formally committed to issuing no more such licences as part of the Beyond Oil and Gas Alliance. Meanwhile, the cartel of some of the largest oil- and gas-producing countries, OPEC, expects global oil demand to increase from 97 million barrels per day in 2021 to 110 million barrels in 2045, mainly owing to the increased energy needs of low- and middle-income countries.

There is, however, one important caveat to this scenario: it's not clear whether the OPEC estimates, which were published in October 2022, account for the latest landmark climate bill signed into law by US President Joe Biden last August. The United States is the world's leading producer and consumer of oil, representing around one-fifth of the global total. Biden's deal, together with earlier legislation from 2021, will invest more than US$500 billion in climate and energy technologies over the next decade. The law will release significant funding and tax credits to hasten the shift to clean-energy generation, decarbonize industry and promote electrification in transport. It also includes funding for improving households' energy efficiency, and for training and development programmes to help former fossil-fuel workers find jobs. These are rare — if not unprecedented — actions from a major oil producer.



[Cement and steel — nine steps to net zero]

However, the US legislation leaves the door open to new oil and gas leasing by the federal government, which the Biden administration had wanted to shut down. The provision, which reflects concerns about energy prices and the loss of fossil-fuel jobs, came courtesy of Joe Manchin, a Democratic senator from coal country in West Virginia, whose vote was crucial for the law's approval by Congress.

Government action to regulate existing markets and foster new ones can work to bend the climate curve. Take electric vehicles (EVs). A decade ago, global sales stood at just 120,000 a year. By 2021, annual sales worldwide had reached 6.6 million — double the number in 2020 — according to the IEA, and 2 million were sold in the first quarter of 2022 alone, 75% more than in the same period of 2021.

There are many reasons for this increase, the IEA says in its Global EV Outlook 2022 report (see go.nature.com/3vyyrsu), including subsidies and the provision of accessible charging. But, the report says, "sustained policy support is the main pillar". This includes legislation to phase out internal combustion engines in new cars by specific dates, giving industry little choice but to shift. So far, more than 20 countries have said that they will do this over the next 10 to 30 years, including several middle-income nations.

The question that researchers, industry leaders and policymakers need to ask this year is what kinds of policy (or regulation) could similarly bend the curve in broader energy markets; in freight transport, aviation and shipping; in industrial processes such as cement and steel manufacture; and in food production and farming. Do countries need to start setting end dates? Can the EV experience be repeated in other sectors? And can researchers model the consequences of such actions, both good and bad?

The COPs have created momentum and pressure for coordinated action, but the influence of oil and gas interests will continue to limit their ambition. As a group, lobbyists representing the fossil-fuel industries outnumbered almost all national delegations to COP27. That's why more countries must commit to ending new oil and gas development, either independently or collectively, through partnerships such as the Beyond Oil and Gas Alliance.

It's clear that technologies for decarbonization are either ready or well on their way to being ready. The same cannot be said for their implementation. Researchers, campaigners, industry and, above all, decision makers must figure out which policy steps are needed, and which will work, to unlock the future we all want to see.

Nature **613**, 216 (2023)

doi: https://doi.org/10.1038/d41586-023-00039-8

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00039-8

- WORLD VIEW
- 11 January 2023
- Correction 13 January 2023

Why NASA and federal agencies are declaring this the Year of Open Science



Here's how NASA is incentivizing open science, and how you can too.

- Chelle Gentemann

I'm thrilled to be the Transform to Open Science lead for NASA, which has a 60-year legacy of pushing the limits of how science is used to understand the Universe, planetary systems and life on Earth. Much of NASA's success can be attributed to a culture of openness for the public good. Since the 1990s, the agency has been a leading advocate for full and open access to data and algorithms.

That culture is needed now more than ever. Humanity is facing many intersecting challenges, from the COVID-19 pandemic to climate change and food and water insecurity. To combat them, we must find breakthroughs faster, increase interdisciplinary expertise and improve how we translate research findings into action. This will require a fundamental shift: from simply sharing results in journal articles to collaborating openly, publishing reproducible results and implementing full inclusivity and transparency.

To catalyse this shift, on 11 January the US White House — joined by 10 federal agencies, a coalition of more than 85 universities, and other organizations — declared 2023 to be the Year of Open Science.



Many researchers say they'll share data — but don't

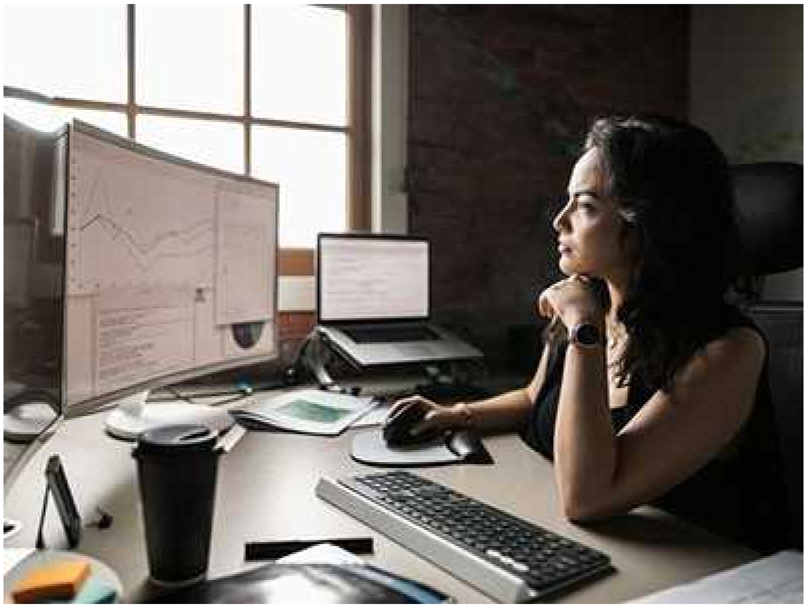
I'm a computational oceanographer and have worked on NASA missions since 1993. I became an enthusiastic open-science advocate after I attended a 2018 workshop at which we reproduced a set of famous figures that analysed satellite data to reveal rising sea levels (see go.nature.com/3go8i). Using a cloud-based data set and open-source Python tools, we did in a couple of minutes what probably originally took months. It was a moment when I could see the future and just leapt because it was so beautiful.

At my next workshop, I realized that open science isn't just about tools. Open-science innovation is being driven by a global community with diverse perspectives. The scientific questions are more interesting and

nuanced, the solutions better. That's what really hooked me. Active inclusion of diverse groups of people is a crucial element of truly open science. I've been on the outside looking in, and it doesn't feel great. But I finally felt like I was situated to do something to remove many barriers to participation.

In May 2021, I sent a one-page call to action for a Year of Open Science to NASA's chief science data officer, and received immediate support. NASA headquarters formed a team to develop the concept. We talked to as many people as possible to learn their motivations, concerns and future needs related to open science. After a year of such discussions, we had a path forward. In April 2022, I started an assignment at NASA to lead the 5-year, US$40-million-dollar Transform to Open Science mission, which will be kicked off with the year of open science.

Science isn't happening just at NASA. But agency partnerships take years to negotiate. I needed a side door. Knowing this, NASA invited me to become one of their representatives to the National Science and Technology Council Subcommittee on Open Science, which coordinates and advances open science across federal agencies. There, I helped to catalyse the Year of Open Science. First, we agreed on a definition: open science is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity. Next, we set four goals for each agency involved in the Year of Open Science: to develop a strategic plan for open science; improve the transparency and equity of reviews; account for open-science activities in evaluations; and engage under-represented communities in the advancement of open science.



Data sharing and how it can benefit your scientific career

Anyone can apply these four goals in their research team, department or organization. So, let's join together. A useful starting point for developing a plan is to look at recommendations from the National Academies and other global organizations. Think about ways to be more open and equitable in reviews — for tenure, promotion, funding and projects. Share the steps that your organization has taken to reduce biases and how you measure their impact. Develop plans or pilot a programme to reward open-science activities in evaluations. Incentives are powerful: they will build a culture of valuing open science.

During the Year of Open Science, NASA is hosting free in-person and virtual workshops to teach basic open-science skills: how to participate, how to write a data-management or software-management plan, and how to share data, software and results. Enrolment has already started (see go.nature.com/3gh8j). By May 2023, there will be five modules, and anyone who completes all five will receive a NASA Open Science Certification to add to their CV and funding proposals. Workshops are planned for 12 scientific conferences in 2023, reaching many of the 100,000 expected attendees.

Transparency requires breaking down barriers that historically have left many out of the scientific community. Last year, NASA committed $20

million per year to advance open science, beginning in 2023. NASA will collaborate with several US minority-serving institutions to ensure that their students and faculty members are true collaborators in the open-science conversation. Our intention is to use in-person and virtual experiences to bridge the visible and invisible barriers of entry into science.

As your organization develops its plans, ensure that under-represented communities are involved from the start. We can’t reach the equitable, open scientific future we need without a diversity of voices. To change everything, we need everyone.

Nature **613**, 217 (2023)

doi: https://doi.org/10.1038/d41586-023-00019-y

Updates & Corrections

- **Correction 13 January 2023**: This World View used the wrong name for the White House’s action to advance equitable science. It is the Year of Open Science, not the Federal Year of Open Science.

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00019-y

News in Focus

- **Are quantum computers about to break online privacy?** [06 Jan 2023]
  • A new algorithm is probably not efficient enough to crack current encryption keys — but that's no reason for complacency, researchers say.
- **Coronavirus variant XBB.1.5 rises in the United States — is it a global threat?** [09 Jan 2023]
  • Prevalence of a new subvariant of Omicron is increasing, but whether it will cause a big surge in infections or hospitalizations isn't clear.
- **COVID drug Paxlovid was hailed as a game-changer. What happened?** [03 Jan 2023]
  • Insufficient investment and fears about rebound and side effects are driving down use of a lifesaving antiviral.
- **'Disruptive' science has declined — and no one knows why** [04 Jan 2023]
  • The proportion of publications that send a field in a new direction has plummeted over the past half-century.
- **China's Thousand Talents Plan to entice researchers home boosted their output** [05 Jan 2023]
  • Analysis of Young Thousand Talents Plan comes amid mounting competition and suspicion between the United States and China.
- **FDA approves Alzheimer's drug lecanemab amid safety concerns** [07 Jan 2023]
  • Reports of deaths potentially linked to the treatment have cast a shadow on what many hail as a landmark approval.
- **Climate-change content shrinks in US university textbooks** [21 Dec 2022]
  • Sections on climate change have become shorter and moved farther back in biology textbooks since the 2000s.
- **How India's caste system limits diversity in science — in six charts** [11 Jan 2023]

• Data show how privileged groups still dominate many of the country’s elite research institutes.

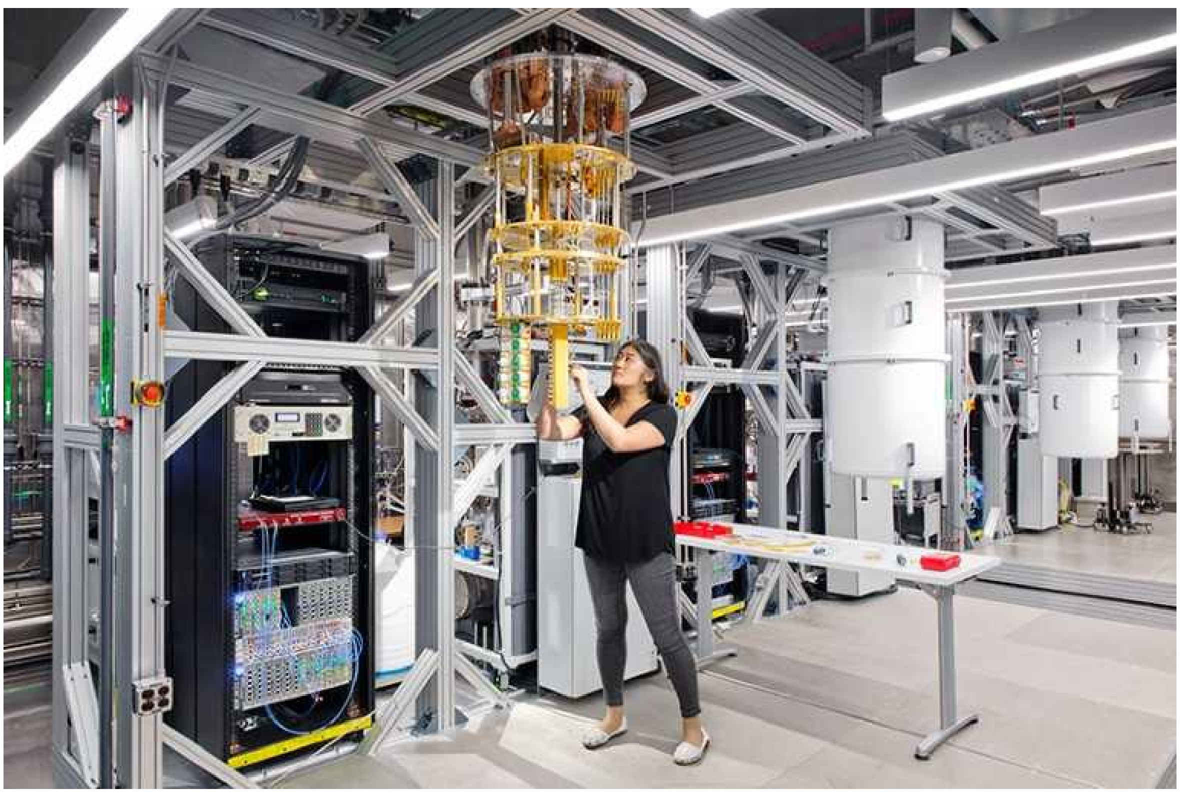
- **Colonoscopies save lives. Why did a trial suggest they might not?** [10 Jan 2023]
• A major clinical study raised questions about one of the most celebrated cancer-screening procedures available, but a close look at the data tells a different story.

- NEWS
- 06 January 2023

Are quantum computers about to break online privacy?

A new algorithm is probably not efficient enough to crack current encryption keys — but that's no reason for complacency, researchers say.

- Davide Castelvecchi



A quantum computer at IBM's Thomas J. Watson Research Center in New York.Credit: Connie Zhou for IBM

A team of researchers in China has unveiled a technique that — theoretically — could crack the most common methods used to ensure digital privacy, using a rudimentary quantum computer.

The technique worked in a small-scale demonstration, the researchers report, but other specialists are sceptical that the procedure could be scaled up to beat ordinary computers at the task. Still, they warn that the paper, posted late last month on the arXiv repository[1], is a reminder of the vulnerability of online privacy.

Quantum computers are known to be a potential threat to current encryption systems, but the technology is still in its infancy. Researchers typically estimate that it will be many years until quantum computers can crack cryptographic keys — the strings of characters used in an encryption algorithm to protect data — faster than ordinary computers.

Researchers realized in the 1990s that quantum computers could exploit peculiarities of physics to perform tasks that seem to be beyond the reach of 'classical' computers. Peter Shor, a mathematician who is now at the Massachusetts Institute of Technology in Cambridge, showed[2] in 1994 how to apply the phenomena of quantum superposition — which describes the ability of atomic-sized objects to exist in a combination of multiple states at the same time — and quantum interference, which is analogous to how waves on a pond can add to each other or cancel each other out , to factoring integer numbers into primes, the integers that cannot be further divided without a remainder.

Shor's algorithm would make a quantum computer exponentially faster than a classical one at cracking an encryption system based on large prime numbers — called Rivest–Shamir–Adleman, or RSA, after the initials of its inventors — as well as some other popular cryptography techniques, which currently protect online privacy and security. But implementing Shor's technique would require a quantum computer much larger than the prototypes that are available. The size of a quantum computer is measured in quantum bits, or qubits. Researchers say it might take one million or more qubits to crack RSA. The largest quantum machine available today — the Osprey chip, announced in November by IBM — has 433 qubits.

A fresh approach

Shijie Wei at the Beijing Academy of Quantum Information Sciences and collaborators took a different route to beat RSA, based not on Shor's but on Schnorr's algorithm — a process for factoring integer numbers devised by mathematician Claus Schnorr at Goethe University in Frankfurt, Germany, also in the 1990s. Schnorr's algorithm was designed to run on a classical computer, but Wei's team implemented part of the process on a quantum computer, using a procedure called the quantum approximate optimization algorithm, or QAOA.

In the paper, which has not yet been peer reviewed, the authors claim that their algorithm could break strong RSA keys — numbers with more than 600 decimal digits — using just 372 qubits. In an e-mail to *Nature* on behalf of all the authors, Guilu Long, a physicist at Tsinghua University in China, cautioned that having many qubits is not enough, and that current quantum machines are still too error-prone to do such a large computation successfully. "Simply increasing the qubit number without reducing the error rate does not help."

Chao-Yang Lu, a physicist who builds quantum computers at the University of Science and Technology of China in Hefei and who was not involved in the project, says that running the QAOA algorithm on such a small machine would require each of the 372 qubits to work without errors 99.9999% of the time. State-of-the-art qubits have barely reached 99.9% accuracy.

The team demonstrated the technique on a 10-qubit quantum computer to factor the more-manageable, 15-digit number 261,980,999,226,229. (It splits into two primes, as 15,538,213 × 16,860,433.) The researchers say this is the largest number yet to have been factored with the aid of a quantum computer — although it is much smaller than the encryption keys used by modern web browsers.

Controversial paper

The trouble is, no one knows whether the QAOA makes factoring large numbers faster than just running Schnorr's classical algorithm on a laptop.

“It should be pointed out that the quantum speedup of the algorithm is unclear,” write the authors. In other words, although Shor’s algorithm is guaranteed to break encryption efficiently when (and if) a large-enough quantum computer becomes available, the optimization-based technique could run on a much smaller machine, but it might never finish the task.

Michele Mosca, a mathematician at the University of Waterloo in Canada, also points out that the QAOA is not the first quantum algorithm known to be able to factor whole numbers using a small number of qubits. He and his collaborators described[3] one in 2017. So researchers already knew that there is nothing fundamental that requires quantum computers to be very large to factor numbers.

Other researchers have complained that, although the latest paper could be correct, the caveat regarding speed comes only at the very end of it. “All told, this is one of the most misleading quantum computing papers I’ve seen in 25 years,” blogged quantum-computing theorist Scott Aaronson at the University of Texas at Austin.

In his e-mail, Long says that he and his collaborators plan to change the paper and will move the caveat higher up. “We welcome the peer review and the communication with scientists around the world,” the statement added.

Even if the Schnorr-based technique won’t break the Internet, quantum computers could eventually do so by running Shor’s algorithm. Security researchers have been busy developing a number of alternative cryptographic systems that are seen as less likely to succumb to a quantum attack, called post-quantum or quantum-safe. But researchers might also discover better quantum algorithms in the future that can beat these systems, with calamitous consequences.

“Confidence in digital infrastructures would collapse,” says Mosca. “We’d suddenly switch from managing the quantum-safe migration through technology lifecycle management to crisis management,” he adds. “It won’t be pretty any way you slice it.”

Nature **613**, 221-222 (2023)

doi: https://doi.org/10.1038/d41586-023-00017-0

References

1. Yan, B. *et al*. Preprint at https://arxiv.org/abs/2212.12372 (2022).
2. Shor, P. W. *Phys. Rev. A* **52**, R2493–R2496 (1995).
3. Bernstein, D. J., Biasse, J.-F. & Mosca, M. in *Post-Quantum Cryptography* Vol. 10346 (eds Lange, T. & Takagi, T.) 330–346 (Springer, 2017).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00017-0

- NEWS
- 09 January 2023

Coronavirus variant XBB.1.5 rises in the United States — is it a global threat?

Prevalence of a new subvariant of Omicron is increasing, but whether it will cause a big surge in infections or hospitalizations isn’t clear.

- Ewen Callaway



A driver is tested for COVID-19.Credit: US Army/ZUMA Press Wire Service/Shutterstock

New year, new variant. Just as scientists were getting to grips with the alphabet soup of SARS-CoV-2 variants circulating globally — your BQ.1.1, CH.1.1 and BF.7 — one lineage seems to be rising to the top, thanks to a peculiar new mutation.

The XBB.1.5 subvariant now makes up around 28% of US COVID-19 cases, according to projections from the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, and its prevalence is on the rise globally. In the Northeastern United States, it seems to have rapidly out-competed the menagerie of other immunity-dodging variants that were expected to circulate alongside one another this winter.

"It's almost certainly going to dominate in the world. I cannot find a single competitor now. Everything else is incomparable," says Yunlong Cao, an immunologist at Peking University in Beijing whose team is studying the properties of XBB.1.5 in the laboratory.

Scientists caution that XBB.1.5's impact, in the United States and beyond, is still far from clear. The variant might not cause a big surge in infections or hospitalizations in many countries, thanks to high immunity built up from exposure to earlier waves of COVID-19 and vaccinations, particularly recent boosters for individuals most at risk of severe disease.

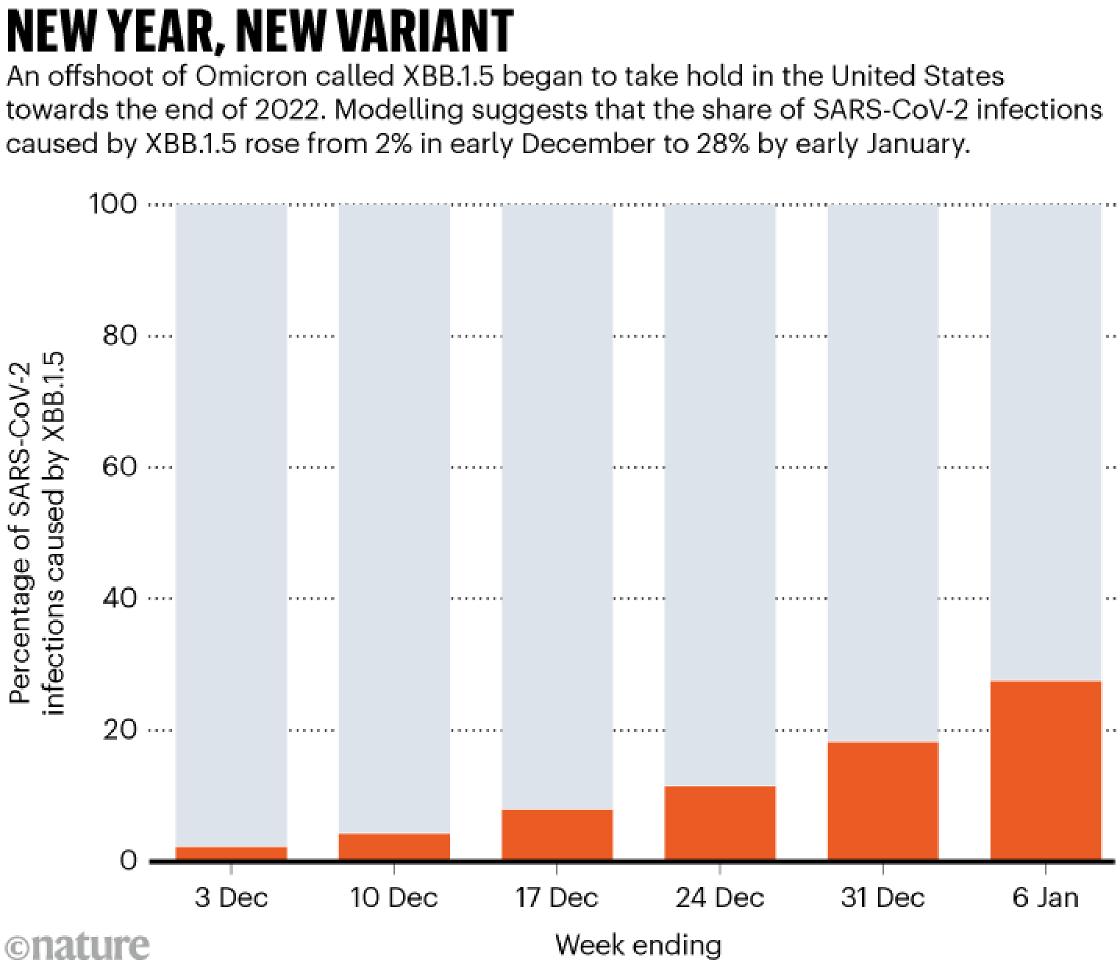
However, even if XBB.1.5 does not cause big COVID-19 waves, it will be important to track the lineage closely, researchers say. The subvariant bears a rarely seen mutation that might enhance its infectivity — and create an opportunity for further evolutionary gains.

Great-grandchild of Omicron

As its name suggests, XBB.1.5 is an offshoot of a SARS-CoV-2 variant called XBB. That lineage is a recombinant of two descendants of the BA.2 lineage that began spiking in early 2022; BA.2 itself is an offshoot of Omicron. XBB's spike protein has a suite of mutations that boost the variant's ability to evade antibodies. This has helped XBB to become common over the past few months, particularly in Asia, where it caused a surge in cases in Singapore.

Variant-watchers noticed XBB.1.5 in late 2022, thanks to a rarely seen amino-acid change, called F486P, in the spike protein. Experiments from Cao's lab suggest that the mutation improves the variant's ability to attach to the human ACE2 receptor, which SARS-CoV-2 uses to invade cells[1]. Importantly, the mutation doesn't seem to erode XBB's prowess at eluding antibodies. The results were posted to the bioRxiv preprint server on 5 January and have not yet been peer reviewed. "XBB really sucks at ACE2 binding," says Cao, and the F486P change present in XBB.1.5 helps to surmount that shortcoming.

The relationship between a variant's ability to attach to ACE2 and its transmissibility isn't fully clear, says Jesse Bloom, an evolutionary virologist at the Fred Hutchinson Cancer Center in Seattle, Washington. But for XBB.1.5, "F486P seems to have given it another boost, which is enabling the virus to spread", he says.



Source: US Centers for Disease Control and Prevention

The CDC estimates that XBB.1.5 is the second most common variant in the United States, comprising 28% of cases nationally, and upwards of 70% in the northeast (see ‘New year, new variant‘). Moritz Gerstung, a computational biologist at the German Cancer Research Centre in Heidelberg, estimates that cases of the variant are doubling roughly every week in the United States, and a bit more slowly in other countries where the variant has appeared. That’s comparable to the rate at which the BQ.1 and BQ.1.1 variants grew in September 2022, but slower than earlier Omicron waves. “XBB.1.5’s spread is still impressively fast,” Gerstung says.

What’s not clear is whether such growth will be sustained or whether the variant will drive up infections significantly, Gerstung adds. BQ.1 and BQ.1.1 looked set to drive sizeable waves, only to run out of steam in

Europe and North America. If the same thing happens with XBB.1.5, the lineage could wind up silently replacing other variants in some countries without causing a big rise in cases.

Big-city variant

Jennifer Surtees, a biochemist at the University at Buffalo in New York, wonders whether researchers are overestimating XBB.1.5's growth in the Northeastern United States. The variant has become more common in the western New York sequences her team handles, but she hasn't yet noticed the meteoric rise in XBB.1.5 genomes that labs in New York City are recording.

Gauging XBB.1.5's impact might not be straightforward, owing to the drop-off in testing for COVID-19, Surtees adds. "I think that we are truly flying blind right now. We have no idea how many cases are really out there."

Tulio de Oliveira, a bioinformatician at Stellenbosch University in South Africa, thinks researchers should look at hospital cases and other measures of disease severity to best measure XBB.1.5's impact. Factors such as a cold snap in the Northeastern United States and holiday gatherings could partly explain the variant's apparent surge, he says. "I think that many scientists are jumping to conclusions and predictions very early and with very little data."

Evasion expert

One thing that researchers can agree on is that XBB.1.5, like its predecessor XBB, is a master of immune evasion. It carries numerous spike mutations that blunt the potency of antibodies raised by vaccination and infection — including earlier Omicron strains. Bivalent vaccines boost levels of antibodies capable of blocking XBB infection (and probably XBB.1.5) in lab tests[2,3], but not by much, notes Cao.

Throughout 2022, researchers including Cao watched Omicron lineages pick up a succession of antibody-evading mutations in the viral spike protein that allowed new lineages to overcome immunity gained from vaccines and previous waves. XBB.1.5 is vastly more transmissible than other circulating

variants thanks to the addition of the F486P mutation, so there is currently little evolutionary pressure on the lineage to change further, says Cao.

But as global immunity to the subvariant builds, XBB.1.5 won't stand still, he says. "We are going to see a lot of new mutations that we have never seen before."

Nature **613**, 222-223 (2023)

doi: https://doi.org/10.1038/d41586-023-00014-3

References

1. Yue, C. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2023.01.03.522427 (2023).
2. Zou, J. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.11.17.516898 (2022).
3. Davis-Gardner, M. E. *et al. N. Engl. J. Med*. https://doi.org/10.1056/NEJMc2214293 (2022).

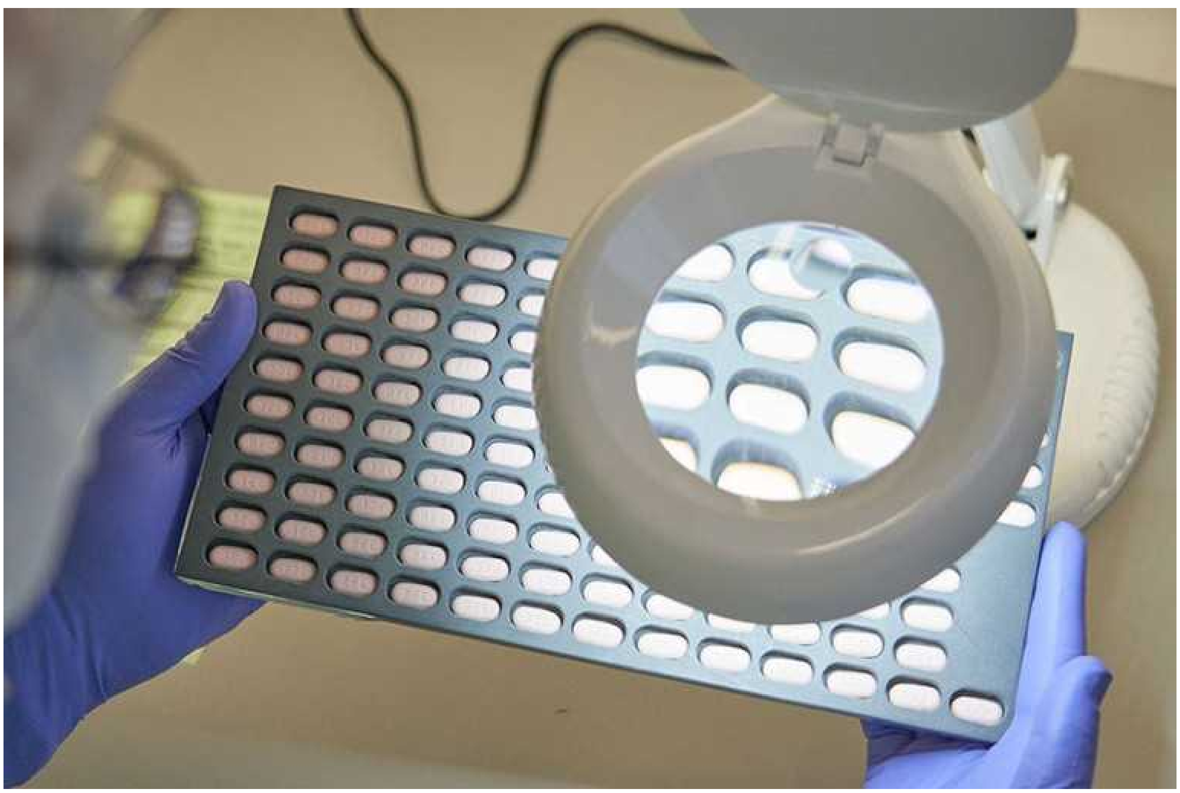
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00014-3

- NEWS
- 03 January 2023

COVID drug Paxlovid was hailed as a game-changer. What happened?

Insufficient investment and fears about rebound and side effects are driving down use of a lifesaving antiviral.

- Max Kozlov



Tablets of Paxlovid, which reduces risk of severe COVID-19, are inspected by a laboratory technician.Credit: Thomas Hansmann/AP/Shutterstock

When clinical trial data for the antiviral drug Paxlovid emerged in late 2021, physicians hailed its astonishing efficacy — a reduction of nearly 90% in the risk of severe COVID-19. But more than a year later, COVID-19 remains a leading cause of death in many countries, and not only in low-income nations where the drug is in short supply. In the United States, for example, hundreds of people still die from COVID-19 each day.

Researchers say that the drug's rollout has been hampered by worries about 'rebound' (the mysterious return of symptoms or detectable virus days after a person starts to feel better) and side effects — as well as by declining concern about the risk of COVID-19. Inadequate funding for distribution, the drug's high price tag and the need for it be taken soon after infection have also slowed its uptake. As a result, physicians have prescribed the drug in only about 0.5% of new COVID-19 cases in the United Kingdom, and in about 13% in the United States, according to a report by the health-analytics firm Airfinity, based in London, UK. Even doctors have reported serious difficulties in helping their family members to obtain Paxlovid[1].

Sentiment against the drug has persisted even as regulators globally have rescinded authorizations for monoclonal antibodies against COVID-19, leaving Paxlovid as one of the only tools to prevent death in high-risk individuals, says Davey Smith, an infectious-disease physician at the University of California, San Diego. "It's a game-changer drug that has good efficacy, even in the setting of Omicron," says Smith. "But rebound has been tagged as a reason not to take the drug, which is a shame."

Premium protection

Paxlovid is a combination of the oral antiviral drugs nirmatrelvir and ritonavir. It reduced the risk of hospitalization or death by 89% in high-risk individuals who took the drug within three days of experiencing symptoms, according to a clinical trial[2] sponsored by pharmaceutical company Pfizer, which produces the drug and is based in New York City.

US regulators first authorized Paxlovid in December 2021, and have since loosened restrictions on who can prescribe it in an effort to make it more broadly available. But health officials lament that the drug has not been

deployed to the extent they had expected: about 10 million Paxlovid courses have been delivered to the United States, but only about 6.7 million have been used.

This underuse stems in part from misinformation and misconceptions about the drug, says Daniel Griffin, an infectious-disease physician at Columbia University in New York City.



Science misinformation alarms Francis Collins as he leaves top NIH job

For example, people perceive COVID-19 as less risky than they did earlier in the pandemic, Griffin says, making them less likely to seek treatment and physicians less likely to prescribe medications promptly. Paxlovid works by inhibiting viral replication, which mostly occurs early in the disease course. This means that the drug must be taken within the first five days of symptom onset, leaving a narrow window of time for people to receive treatment. That does not mesh well with "this mentality of 'Let's wait and see how you do'," among doctors treating people with COVID-19, says Griffin.

To add to the confusion, worries have swirled about post-Paxlovid 'rebound', fuelled by high-profile cases in figures such as US President Joe Biden and and former US infectious-disease chief Anthony Fauci. Such publicity has had a chilling effect on the number of people seeking Paxlovid, says Smith.

But researchers have found that rebound often occurs even in people who don't take Paxlovid[3]. Precise estimates for rebound incidence vary, depending on the population studied and the definition of 'rebound'. But regardless of whether people take Paxlovid, Smith says, it's common for them to experience either viral rebound — in which people test positive again — or symptom rebound, but not both at the same time[4]. Smith says that symptom rebound tends to be very mild, and is still far preferable to hospitalization or death.

A bitter pill

In addition, Paxlovid can't be taken with many other drugs and, in some people, makes certain foods taste bitter or metallic, both of which reinforce the perception that it's toxic, Smith says. Treatment requires taking several pills twice a day for five consecutive days — which isn't the easiest regimen to adhere to, Smith adds.

Such concerns have led health officials to point to hesitancy to explain why use of the drug has fallen short. But there are also systemic reasons, says Anne Sosin, a public-health-policy specialist at Dartmouth College in Hanover, New Hampshire.

Paxlovid relies on a robust COVID-19 testing infrastructure and access to primary-care physicians and pharmacies, she notes. This amplifies pre-existing disparities resulting from race and income. For example, Black and Hispanic populations were about 36% and 30% less likely to be prescribed Paxlovid, respectively, compared with white populations, according to an analysis[5] of almost 700,000 people who sought COVID-19 care across 30 US sites. Fears about hesitancy, says Sosin, provide an excuse to blame individuals rather than policymakers and to deflect "attention away from the system that needs to be in place to deliver the drugs".



Can drugs reduce the risk of long COVID? What scientists know so far

To remedy these disparities, she says she would like to see health officials mount an ‘all hands on deck’ approach to ensure that everybody has equal access to the drug by engaging local communities and expanding access to testing centres. For example, health officials have successfully narrowed similar disparities in the number of people who received their primary COVID-19 vaccination series by bringing vaccines “to people in the areas they live, work and play”.

COVID-19 isn’t going away, says Smith, so it’s important to develop other antiviral drugs. Paxlovid could soon have competition: in November, Japan authorized ensitrelvir, a once-daily antiviral made by the Japanese pharmaceutical company Shionogi, based in Osaka and Hokkaido University in Japan. In July, China authorized the use of an HIV drug to treat COVID-19, but detailed data from a large clinical trial have not been released.

Nature **613**, 224-225 (2023)

doi: https://doi.org/10.1038/d41586-022-04576-6

References

1. Mangurian, C. *JAMA* https://doi.org/10.1001/jama.2022.23155 (2022).

2. Hammond, J. *et al. N. Engl. J. Med.* **386**, 1397–1408 (2022).

3. Anderson, A. S. *et al. N. Engl. J. Med.* **387**, 1047–1049 (2022).

4. Smith, D. M. *et al. JAMA Netw. Open.* **5**, e2238867 (2022).

5. Boehmer, T. K. *et al. MMWR Morb. Mortal. Wkly. Rep.* **71**, 1359–1365 (2022).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-022-04576-6

- NEWS
- 04 January 2023

‘Disruptive’ science has declined — and no one knows why

The proportion of publications that send a field in a new direction has plummeted over the past half-century.

- Max Kozlov



Citation data from 45 million scientific manuscripts reveals a puzzling trend.Credit: Joos Mind/Getty

The number of science and technology research papers published has skyrocketed over the past few decades — but the 'disruptiveness' of those papers has dropped, according to an analysis of how radically papers depart from the previous literature[1].

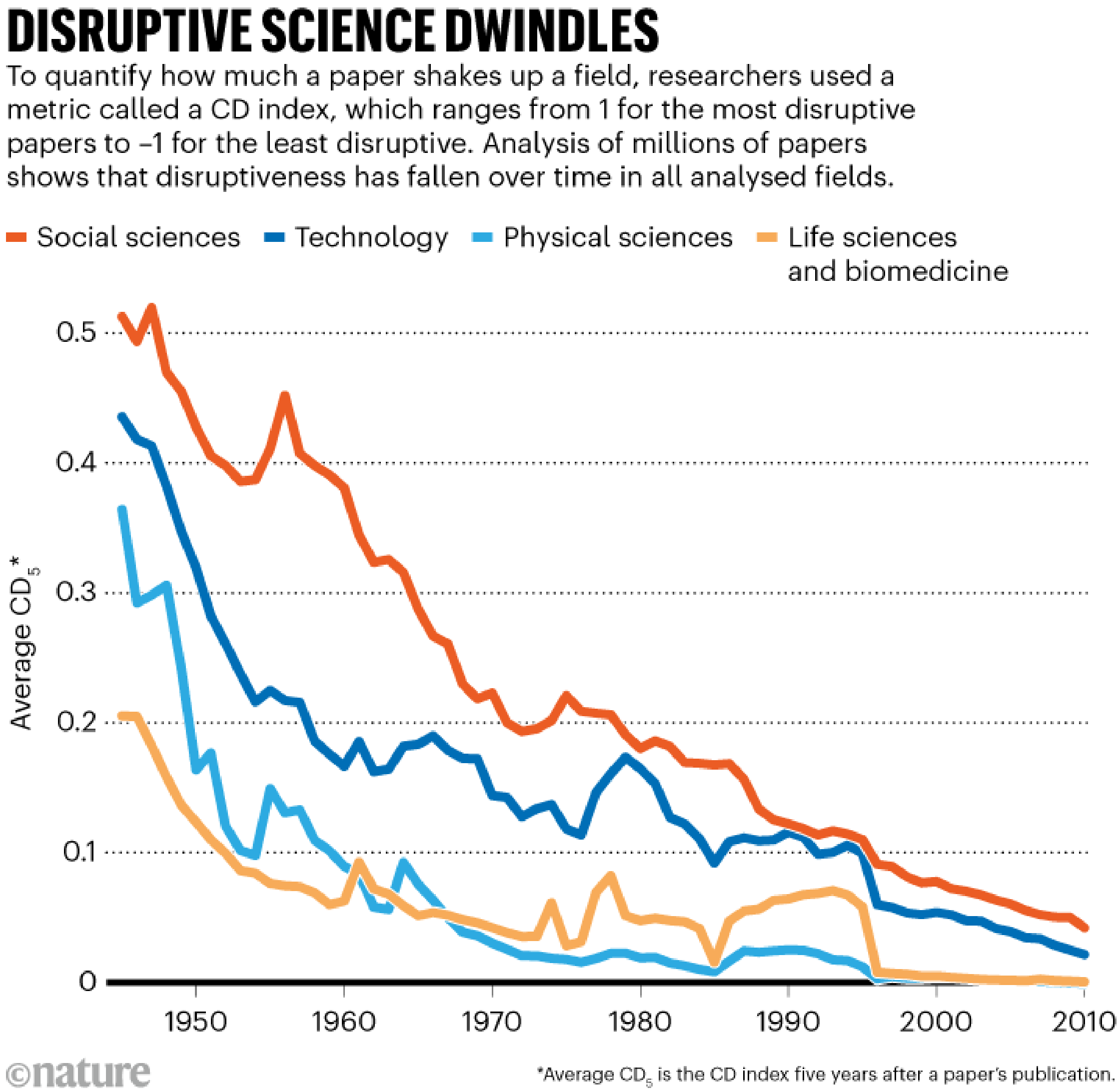
Data from millions of manuscripts show that, compared with mid-twentieth-century research, that done in the 2000s was much more likely to push science forward incrementally than to veer off in a new direction and render previous work obsolete. Analysis of patents from 1976 to 2010 showed the same trend.

"The data suggest something is changing," says Russell Funk, a sociologist at the University of Minnesota in Minneapolis and a co-author of the analysis, which was published on 4 January in *Nature*. "You don't have quite the same intensity of breakthrough discoveries you once had."

Telltale citations

The authors reasoned that if a study was highly disruptive, subsequent research would be less likely to cite the study's references, and instead would cite the study itself. Using the citation data from 45 million manuscripts and 3.9 million patents, the researchers calculated a measure of disruptiveness, called the CD index, in which values ranged from –1 for the least disruptive work to 1 for the most disruptive.

The average CD index declined by more than 90% between 1945 and 2010 for research manuscripts (see 'Disruptive science dwindles'), and by more than 78% from 1980 to 2010 for patents. Disruptiveness declined in all of the analysed research fields and patent types, even when factoring in potential differences in factors such as citation practices.



Source: Ref. 1

The authors also analysed the most common verbs used in manuscripts and found that whereas research in the 1950s was more likely to use words evoking creation or discovery such as 'produce' or 'determine', that done in the 2010s was more likely to refer to incremental progress, using terms such as 'improve' or 'enhance'.

"It's great to see this [phenomenon] documented in such a meticulous manner," says Dashun Wang, a computational social scientist at Northwestern University in Evanston, Illinois, who studies disruptiveness in

science. “They look at this in 100 different ways, and I find it very convincing overall.”

Other research[2] has suggested that scientific innovation has slowed in recent decades, too, says Yian Yin, also a computational social scientist at Northwestern. But this study offers a “new start to a data-driven way to investigate how science changes”, he adds.

Disruptiveness is not inherently good, and incremental science is not necessarily bad, says Wang. The first direct observation of gravitational waves, for example, was both revolutionary and the product of incremental science, he says.

The ideal is a healthy mix of incremental and disruptive research, says John Walsh, a specialist in science and technology policy at the Georgia Institute of Technology in Atlanta. “In a world where we’re concerned with the validity of findings, it might be a good thing to have more replication and reproduction,” he says.

Why the slide?

It is important to understand the reasons for the drastic changes, Walsh says. The trend might stem in part from changes in the scientific enterprise. For example, there are now many more researchers than in the 1940s, which has created a more competitive environment and raised the stakes to publish research and seek patents. That, in turn, has changed the incentives for how researchers go about their work. Large research teams, for example, have become more common, and Wang and his colleagues have found[3] that big teams are more likely to produce incremental than disruptive science.

Finding an explanation for the decline won’t be easy, Walsh says. Although the proportion of disruptive research dropped significantly between 1945 and 2010, the number of highly disruptive studies has remained about the same. The rate of decline is also puzzling: CD indices fell steeply from 1945 to 1970, then more gradually from the late 1990s to 2010. “Whatever explanation you have for disruptiveness dropping off, you need to also make sense of it levelling off” in the 2000s, he says.

Nature **613**, 225 (2023)

doi: https://doi.org/10.1038/d41586-022-04577-5

References

1. Park, M., Leahey, E. & Funk, R. J. *Nature* **613**, 138–144 (2023).
2. Cowen, T. & Southwood, B. Preprint at SSRN http://doi.org/10.2139/ssrn.3822691 (2019).
3. Wu, L., Wang, D. & Evans, J. A. *Nature* **566**, 378–382 (2019).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-022-04577-5

- NEWS
- 05 January 2023
- Correction 10 January 2023

China’s Thousand Talents Plan to entice researchers home boosted their output

Analysis of Young Thousand Talents Plan comes amid mounting competition and suspicion between the United States and China.

- Dyani Lewis



The Thousand Talents Plan aims to make China a global leader in science and technology.Credit: Noel Celis/AFP/Getty

Young Chinese scientists who got their PhDs overseas and returned to China as part of a state-run talent drive published more papers after their return than did their peers who stayed abroad. The productivity bump can be explained by returnees' access to greater funding and an abundant research workforce, according to the authors of an analysis published in *Science*[1]. The findings come as geopolitical competition between the United States and China mounts.



Chinese American scientists uneasy amid crackdown on foreign influence

In 2008, China began enticing expatriate scientists and foreign researchers to relocate to China through its Thousand Talents Plan, which aims to make the nation a global leader in science and technology.

Despite the programme's prestige, government agencies in the United States and elsewhere have regarded it with suspicion, because of its potential to increase the flow of technical know-how to China. In 2018, the administration of former US president Donald Trump launched the China Initiative, to protect US laboratories and businesses from espionage. And the same year, the Chinese government stopped publicly naming Thousand

Talents Plan recipients because of the negative career impact that the association might have.

The latest study focuses on a branch of the programme — the Young Thousand Talents (YTT) Plan — designed to attract early-career researchers to Chinese research posts. The programme offers researchers generous research packages, including an internationally competitive salary and start-up funds: between 2011 and 2017, when records were still publicly available, it offered grants to more than 3,000 researchers.

Researcher calibre

To gauge the programme's success, the authors examined some 300 young Chinese researchers who had got their PhDs abroad, then received YTT grants between 2012 and 2014 and returned to academic positions in China.



China's plan to recruit talented researchers

They found that the grant recipients ranked in the top 15% for publication output in the five years before returning to China, when benchmarked against all early-career scientists in the United States. This suggests that the YTT programme was succeeding at attracting high-calibre scientists, the researchers conclude.

But they also found that scientists who rejected offers to participate in the programme, and stayed abroad, ranked even higher — in the top 10% for research productivity in the five years before they might have returned. These researchers were also more likely to have published as a last author on papers — which usually denotes seniority — in leading journals. Public-policy researcher Kathleen Vogel at Arizona State University in Tempe thinks that the main reason for these findings is that it's hard to persuade top-notch researchers to leave elite institutions, where they already have access to funding and resources.

Productivity boost

During a period of up to seven years after their return to China, YTT scholars experienced a rise in productivity compared with matched researchers who stayed in the United States. Returnees published 27% more papers — equating to approximately one more paper per year on average — including in journals ranked in the top 10% for a field, says co-author Yanbo Wang at the University of Hong Kong.

Returnees were also more likely to be last authors than were those who stayed, suggesting that YTT scholars are better able to forge independent research careers as heads of their own research teams, Wang says.



China hides identities of top scientific recruits amidst growing US scrutiny

Access to funding and research staff seems to be the main driver of the productivity gains: once funding and size of research team are taken into account, the effect on publication output disappears. “Research grants really matter a lot,” Wang adds.

The effects were most pronounced in fields such as biology, medicine and chemistry, which often require expensive equipment and large numbers of people to conduct research. The study authors write that the YTT programme was beneficial to “young expatriates who had the capability but not the funding to run their own labs for independent research”. Policymakers in the United States and the European Union need to address the shortage of funding for early-career researchers in their countries so that scientists there can kick-start independent careers, too, Wang says.

Lili Yang, a higher-education policy researcher at the University of Hong Kong who has also measured productivity in YTT scholars, says that these findings are “super interesting” and reveal a nuanced picture of the roles held by returning scientists.

Mounting tensions

The study comes as friction between the United States and China ratchets up — and this is also influencing researchers’ decisions about where to work, Yang says. “China is witnessing a higher proportion of returnees because of the tensions, which in the short term is good for the development of Chinese research,” she says.



Harvard chemistry chief's arrest over China links shocks researchers

The study makes an important contribution, says geographer Qingfang Wang at the University of California, Riverside, who researches why scientists — including YTT recipients — migrate. But she notes, on the basis of her investigations, that researchers will apply for YTT funding only once they have decided to relocate to China, so it is unclear how effective the programme is at enticing researchers to the country.

And she says that focusing on publication output as a measure of the programme's value overlooks other markers of scientific success, such as the ability to cultivate creative thinking, innovation and interdisciplinary research, which are harder to measure. The aim is not to produce "publication machines", she adds. In 2020, China's ministries of science and technology and education issued guidelines to reduce institutions' reliance on publication as a way of evaluating research performance.

Yang is interested in how the YTT recipients develop their careers — for example, whether they move into leadership roles in which they can influence research culture. But the lack of transparency since 2018 around who receives YTT grants will make it harder to carry out research on later cohorts of the programme — including those who received grants during the

height of the China Initiative and the turbulent pandemic era. “It is a pity,” she says.

Nature **613**, 226-227 (2023)

doi: https://doi.org/10.1038/d41586-023-00012-5

Updates & Corrections

- **Correction 10 January 2023**: An earlier version of this story omitted the affiliation for Yanbo Wang.

References

1. Shi, D., Liu, W. & Wang, Y. *Science* **379**, 62–65 (2023).

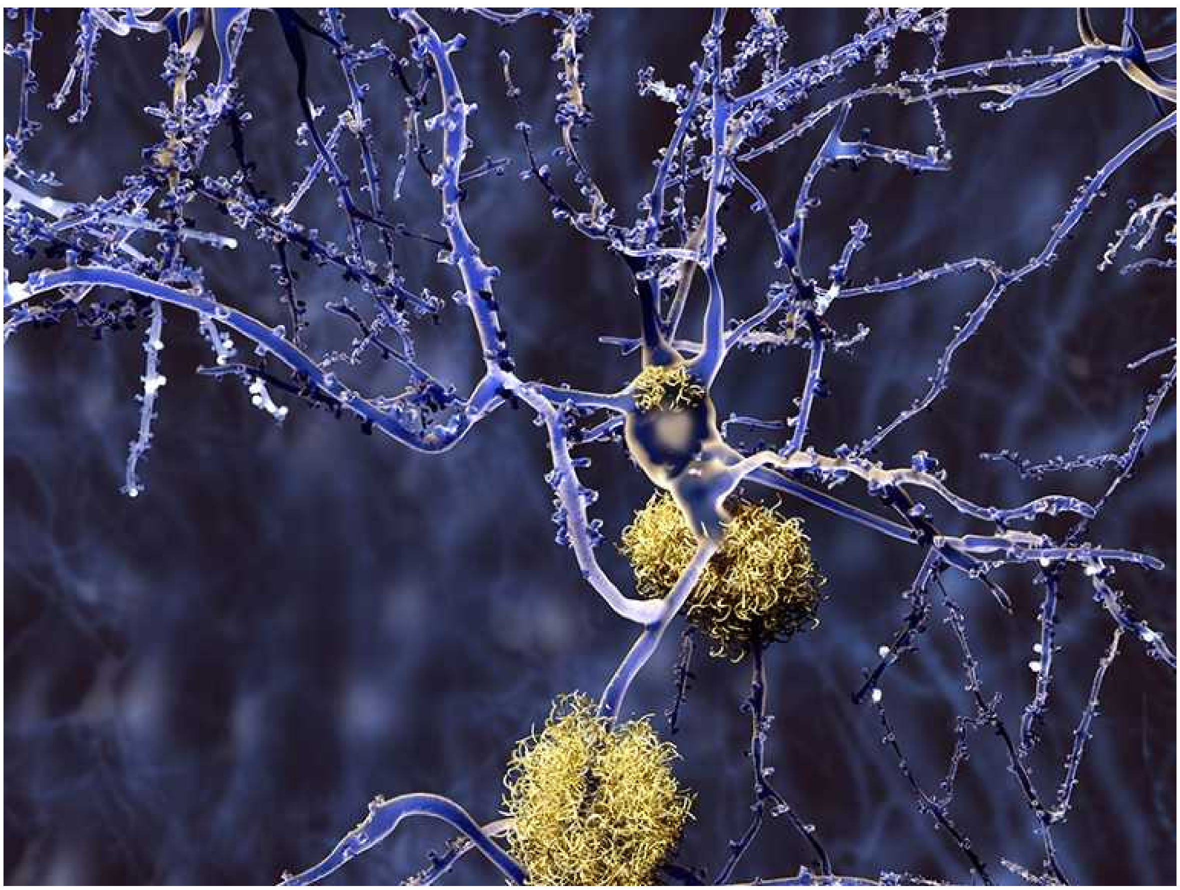
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00012-5

- NEWS
- 07 January 2023
- Correction 10 January 2023

FDA approves Alzheimer's drug lecanemab amid safety concerns

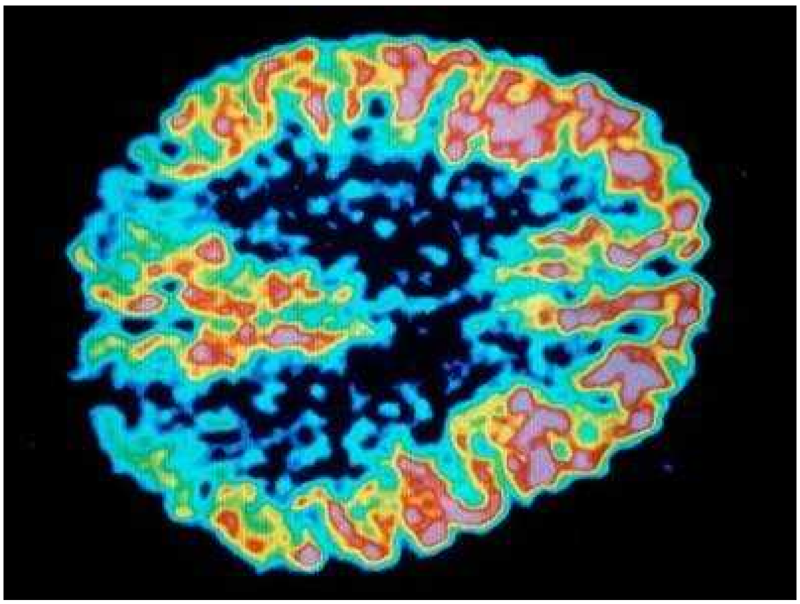
Reports of deaths potentially linked to the treatment have cast a shadow on what many hail as a landmark approval.

- Sara Reardon



Protein clumps called amyloid plaques (gold, among blue neurons in this computer illustration) are a hallmark of Alzheimer's disease.Credit: Juan Gaertner/SPL

The US Food and Drug Administration (FDA) has approved lecanemab, the second-ever treatment for Alzheimer's disease that is intended to tackle the root of the condition and the slow cognitive decline that results. Researchers welcome the decision, but their excitement is clouded by reports of patient deaths and that the FDA acted improperly last year when approving the first drug of this kind.



Heralded Alzheimer's drug works — but safety concerns loom

Because it slows progression of the condition when taken in the early stages, lecanemab will allow individuals to have "more time to participate in daily life and live independently", says Joanne Pike, president and chief executive of the Alzheimer's Association in Chicago, Illinois, in a statement.

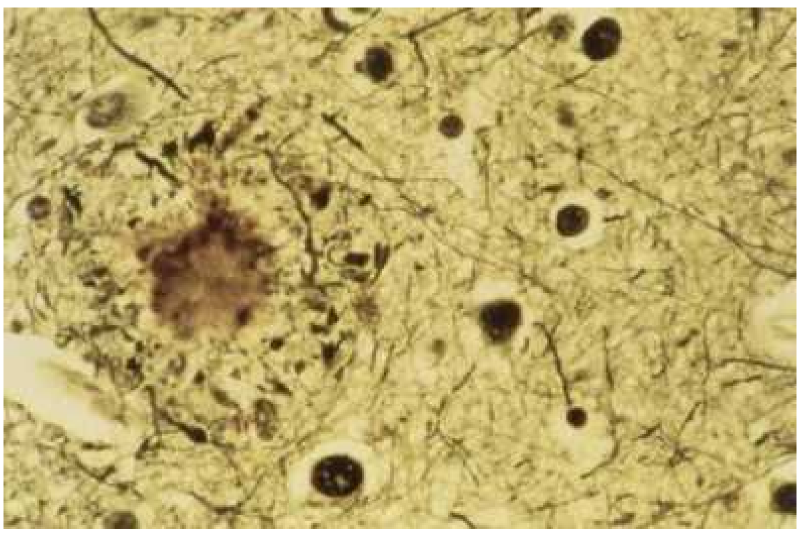
Lecanemab, which will be sold under the brand name Leqembi, is the first Alzheimer's treatment of its kind shown to slow cognitive decline in a robust clinical trial and is the second treatment to be approved in under two years. It is made by the bio-pharmaceutical companies Eisai in Tokyo, and Biogen in Cambridge, Massachusetts. The drug, a monoclonal antibody, is infused into a patient intravenously; it then enters their brain and clears the

amyloid plaques thought to cause cognitive impairment and dementia in Alzheimer's disease.

"The findings provide hope for patients and families, but it is also an important step in the development of successful ways to modify and prevent the clinical onset of the disease," says Eric Reiman, executive director of the Banner Alzheimer's Institute in Phoenix, Arizona.

Unclear impact

Lecanemab was authorized under the FDA's accelerated approval pathway, which is reserved for therapies for diseases with few treatments; it also does not require data from phase III trials. But researchers are hopeful about Biogen and Eisai's phase III data, which were published last November[1]. That trial, conducted on about 1,800 people with early-stage Alzheimer's, found that the antibody slowed cognitive decline by 27% over 18 months of treatment.



Alzheimer's drug slows mental decline in trial — but is it a breakthrough?

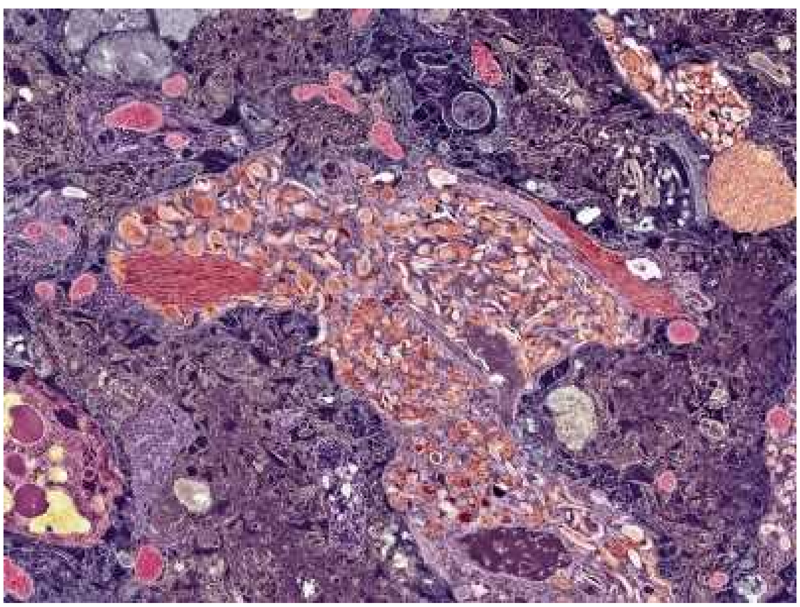
The FDA's decision does not take the phase III study into account — Biogen and Eisai asked for accelerated approval on the basis of phase II data, which they submitted before the latest trial results were announced. The phase II study found that lecanemab decreased brain plaques in 856 participants, but did not assess whether this affected cognitive abilities[2]. This is the same

approval pathway used for its predecessor, aducanumab, another monoclonal antibody made by Biogen and Eisai.

It's unclear what impact that 27% effect will have on the lives of people with Alzheimer's — or whether the effects will persist after 18 months. It could mean, Reiman says, "an extra 6 months of recognizing a loved one's face, or performing a valued activity". But that will require further research, he says. In the meantime, the FDA says, lecanemab is to be used only for people with mild cognitive impairment — the same group tested in the clinical trials.

Treatment risks

Diana Zuckerman, president of the National Center for Health Research, a non-profit organization in Washington DC, isn't sure that the benefit will be worth the risk. "You're dealing with people with mild cognitive impairment who are functioning, and you're putting them at risk" in terms of both their health and their finances, she says. Eisai says that a year of lecanemab will cost US$26,500 in the United States.



Landmark Alzheimer's drug approval confounds research community

In recent months, *Science* and *STAT News* have reported on three people who had been enrolled in the lecanemab phase III study and then died during

the extended phase of the trial, when participants receiving placebos can ask to be given the actual drug. The three people died from complications such as brain bleeding and seizures. According to those reports, researchers think that the patients might have died because of a set of conditions known as amyloid-related imaging abnormalities (ARIA). They suspect that the antibody weakened blood vessels in the brain as it attacked the amyloid plaques lining them. All the patients were taking anticoagulant drugs at the time, which might have worsened the bleeding.

Eisai has said that it is inappropriate to draw conclusions on the basis of individual cases, and that it reported the deaths to the FDA as required. Nevertheless, the FDA's approval mandates that lecanemab include a warning about ARIA, and that physicians monitor for the condition, which the agency says is rarely serious or life-threatening.

The controversy around aducanumab, for which the FDA granted accelerated approval on 7 June 2021, has cast a shadow over lecanemab. Many researchers thought that aducanumab, which also goes by the brand name Aduhelm, did not show a strong signal for cognitive decline. The FDA's scientific advisory panel recommended against approving the antibody in an 8–1 vote, and three panel members resigned after the FDA authorized it anyway. The FDA did not hold a public advisory meeting for lecanemab before approving it.

In aducanumab's shadow

Since aducanumab's approval, Biogen (the drug's lead developer) and the FDA have come under further scrutiny, which has limited the treatment's rollout. A congressional investigation released last month found that the agency had flouted its rules by improperly guiding Biogen through the aducanumab approval process. The report said that the process was "rife with irregularities" and raised "serious concerns about FDA's lapses in protocol and Biogen's disregard of efficacy", but did not impose any penalty on the FDA or Biogen.



More Alzheimer's drugs head for FDA review: what scientists are watching

The US Centers for Medicare and Medicaid Services (CMS) has declined to cover aducanumab under federal insurance plans unless a person is enrolled in a clinical trial, leaving most patients to pay more than $28,000 for a year of treatment. CMS — and several clinics that refused to prescribe the drug — cite its questionable efficacy, drawing the ire of advocacy groups, which say it should be readily available.

Despite the controversies, Reiman hopes that lecanemab's success will open the door to more accelerated approvals. He is currently partnering with pharmaceutical company Eli Lilly, based in Indianapolis, Indiana, to test its similar monoclonal antibody, donanemab, in a phase III trial that he says is showing promising results. The FDA is expected to decide whether to approve donanemab in the coming weeks.

Nature **613**, 227-228 (2023)

doi: https://doi.org/10.1038/d41586-023-00030-3

Updates & Corrections

- **Correction 10 January 2023**: An earlier version of this story indicated that the Alzheimer's Association is based in Washington DC. It is based in Chicago, Illinois.

References

1. Van Dyck, C. H. *et al. N. Engl. J. Med.* **388**, 9–21 (2023).

2. Swanson, C. J. *et al. Alz. Res. Ther.* **13**, 80 (2021).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00030-3

- NEWS
- 21 December 2022

Climate-change content shrinks in US university textbooks

Sections on climate change have become shorter and moved farther back in biology textbooks since the 2000s.

- McKenzie Prillaman



Biology textbooks published in the United States in the 2010s are lighter on climate-change content than those published in the 2000s.Credit: Kristopher Skinner/Bay Area News Group/Alamy

University biology textbooks published in the United States in the 2010s contained less climate-change content than those published in the 2000s, despite the increasing urgency of the climate crisis[1].

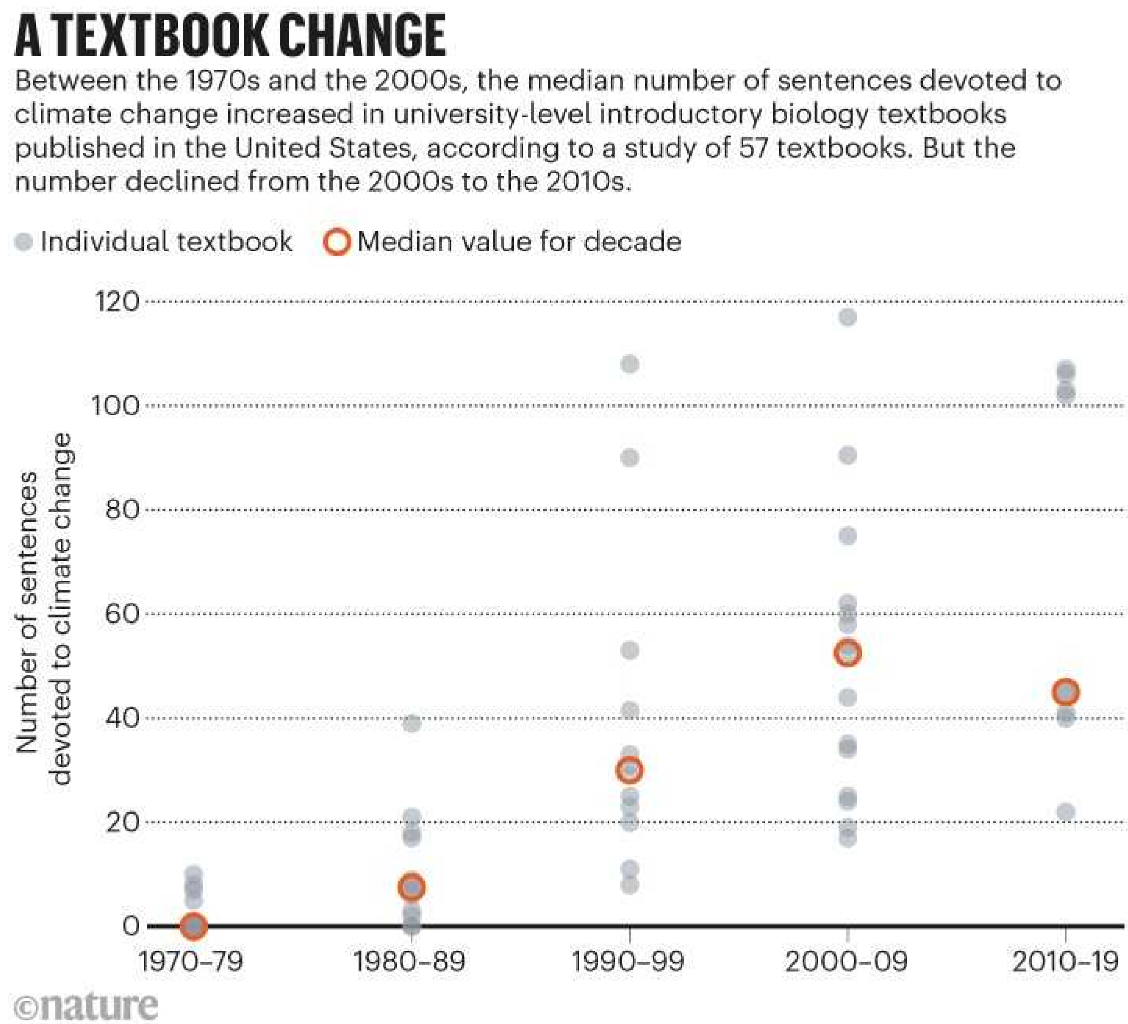
That's not the only shift. In a survey of dozens of widely used introductory biology textbooks from the past 50 years, researchers also found that passages on climate change moved farther back in the books, and that the amount of information about its impacts grew, whereas the space devoted to solutions shrank. The work was published on 21 December in *PLoS ONE*[1].

These textbooks — some of which are also available outside the United States — explain how the changing climate affects a range of complex biological systems, from agriculture to human health. And because they are assigned during entry-level classes, many students reading these books might never take another course that touches on the topic.

"This may actually be the last content on climate change that they get in their formal education," says study co-author Jennifer Landin, a science-education researcher at North Carolina State University in Raleigh. "And if it's three pages, that's not nearly enough."

Reading between the lines

The researchers assessed 57 textbooks published between 1970 and 2019. They counted the sentences in the books' climate-change passages — identified by phrases such as 'global warming' and 'greenhouse gas' — and found that the median number of sentences per passage rose until the 2000s (see 'A textbook change'). It peaked in that decade at 52, but then dropped to 45 in the 2010s. (The authors deemed the median a better value to use than the average because of the wide range in passage length.)



Source: Ref. 1

Furthermore, the median position of sections about climate change moved from the last 15% of pages in the 1970s — when many scientists first became convinced that the planet was warming — to the last 2.5% in the 2010s. Controversial topics are usually placed at the end, Landin says, because “it allows teachers to ‘run out of time’ and then not have to teach them”.

Topics in the climate-change passages also shifted. Since the 1990s, the proportion of sentences devoted to solutions has dwindled, whereas content on the effects of climate change, including extreme weather and worsening diseases, has grown. Landin praises this more holistic portrayal of the

impacts. But with the simultaneous decrease in solutions, she says, the balance of content skews towards hopelessness.

Useful but limited

"It's incredibly useful data," says Mariëlle Hoefnagels, a science-education specialist at the University of Oklahoma in Norman and author of two general biology textbooks, one of which was analysed in the study. She says that the paper's insights concerning climate-change solutions will help her to shape a forthcoming edition of one of her books. However, "books can't continue to grow indefinitely", she adds.



Climate change: Nature readers say their fears are growing

Hoefnagels says that sections on climate change expanded in the 2000s because of a growing understanding of global warming's impacts. But that rise has tapered off now that the fundamentals are well covered in introductory biology textbooks; adding more information would "muscle something else important out of the way", she says.

Landin suggests that the controversy surrounding climate change might be behind its decline in textbooks. But Hoefnagels says she has never felt pressured by her publisher McGraw Hill to tone down 'contentious'

subjects. She says that climate science has remained at the back of books because instructors prefer textbooks that progress from small-scale to large-scale topics, such as from molecular biology to ecology.

Textbook content might not be a reliable guide to how much students are learning about climate change, says Eric Plutzer, a political scientist at Pennsylvania State University in University Park who has surveyed how public-school teachers in the United States cover climate change[2]. Instructors could be spending more time on climate change than textbooks indicate, he adds.

“Textbooks alone give us some valuable information but don’t tell us the whole story,” Plutzer says.

Nature **613**, 228 (2023)

doi: https://doi.org/10.1038/d41586-022-04487-6

References

1. Ansari, R. A. & Landin, J. M. *PLoS ONE* **17**, e0278532 (2022).

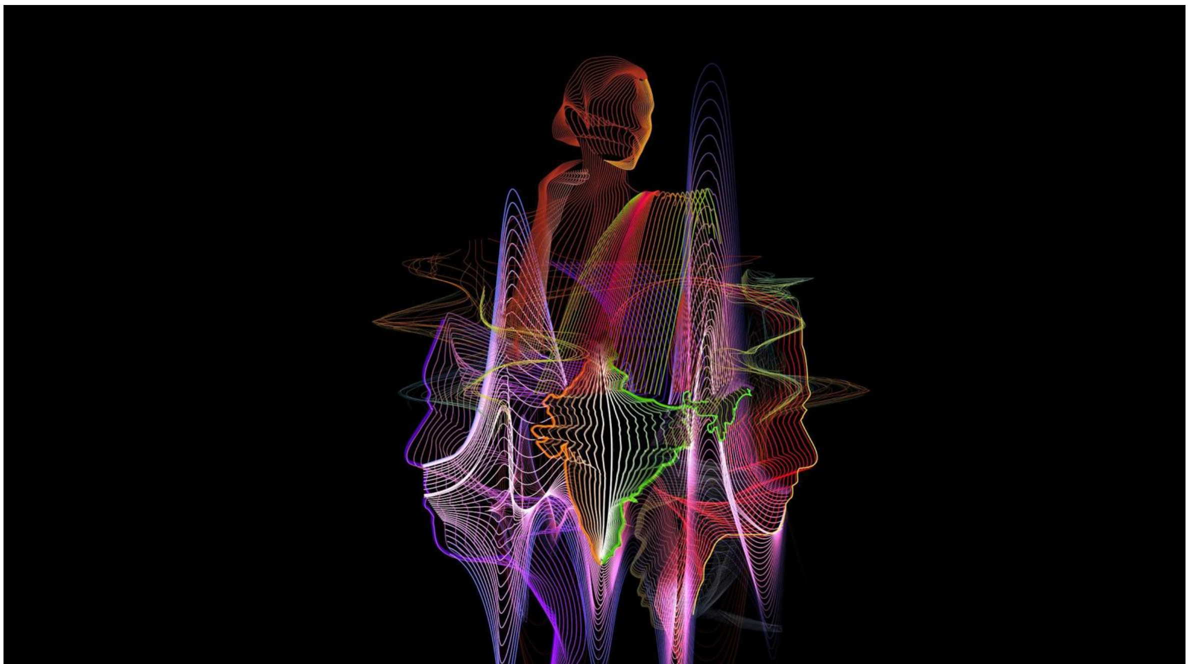
2. Plutzer, E. *et al. Science* **351**, 664–665 (2016).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-022-04487-6

How India’s caste system limits diversity in science — in six charts

Data show how privileged groups still dominate many of the country’s elite research institutes.

By Ankur Paliwal



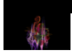


Illustration by Polygon8

Illustration by Polygon8

This article is part of a Nature *series examining data on ethnic or racial diversity in science in different countries. See also: How UK science is failing Black researchers — in nine stark charts.*

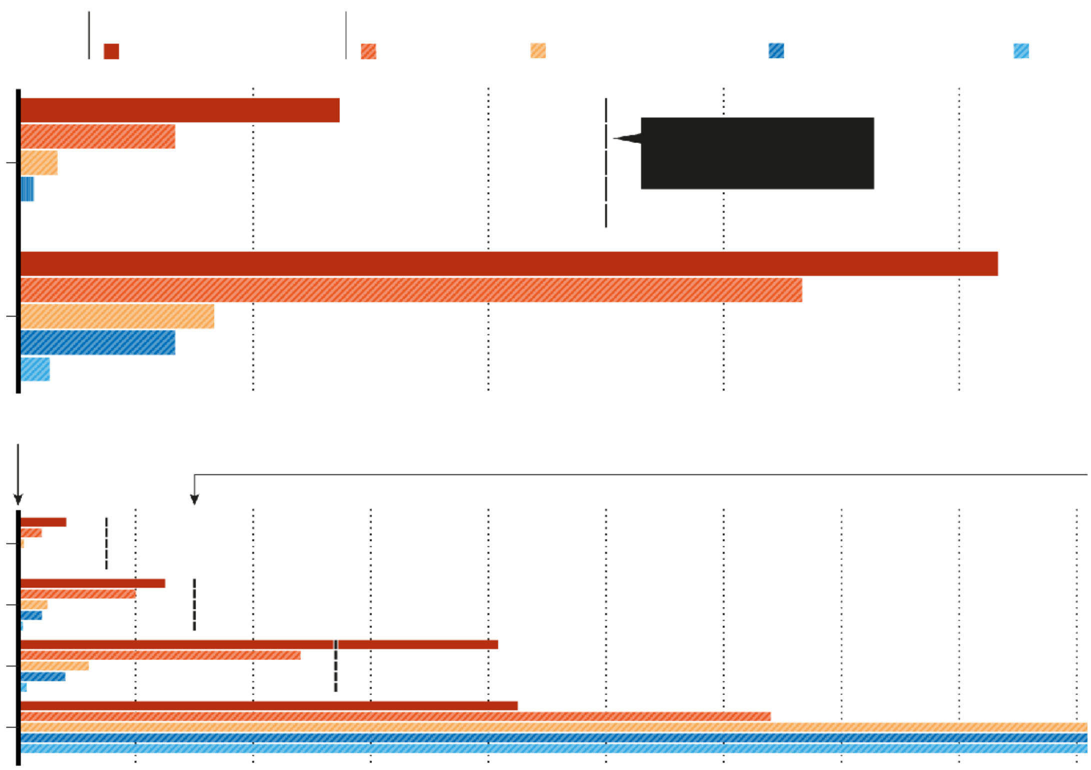
Samadhan is an outlier in his home village in western India. Last year, he became the first person from there to start a science PhD. Samadhan, a student in Maharashtra state, is an Adivasi or indigenous person — a member of one of the most marginalized and poorest communities in India.

For that reason, he doesn’t want to publicize his last name or institution, partly because he fears that doing so would bring his social status to the attention of a wider group of Indian scientists. “They’d know that I am from a lower category and will think that I have progressed because of [the] quota,” he says.

The quota Samadhan refers to is also known as a reservation policy: a form of affirmative action that was written into India’s constitution in 1950. Reservation policies aimed to uplift marginalized communities by allocating quotas for them in public-sector jobs and in education. Mirroring India’s caste system of social hierarchy, the most privileged castes dominated white-collar professions, including roles in science and technology. After many years, the Indian government settled on a 7.5% quota for Adivasis (referred to as ‘Scheduled Tribes’ in official records) and a 15% quota for another marginalized group, the Dalits (referred to in government records as ‘Scheduled

Castes', and formerly known by the dehumanizing term 'untouchables'). These quotas — which apply to almost all Indian research institutes — roughly correspond to these communities' representation in the population, according to the most recent census of 2011.

But the historically privileged castes — the 'General' category in government records — still dominate many of India's elite research institutions. Above the level of PhD students, the representation of Adivasis and Dalits falls off a cliff. Less than 1% of professors come from these communities at the top-ranked institutes among the 23 that together are known as the Indian Institutes of Technology (IITs), according to data provided to *Nature* under right-to-information requests (see 'Diversity at top Indian institutions'; the figures are for 2020, the latest available at time of collection).



Diversity at top Indian institutions

Universities in India are failing to meet government quotas for marginalized communities in academia. The government provides detailed national figures for undergraduate and master's students, but not for academic levels above those; Nature collated figures from right-to-information requests. These data show that the representation of marginalized castes and other groups falls at higher academic levels.

Na

To

As

As

Pro

Ph

Ca

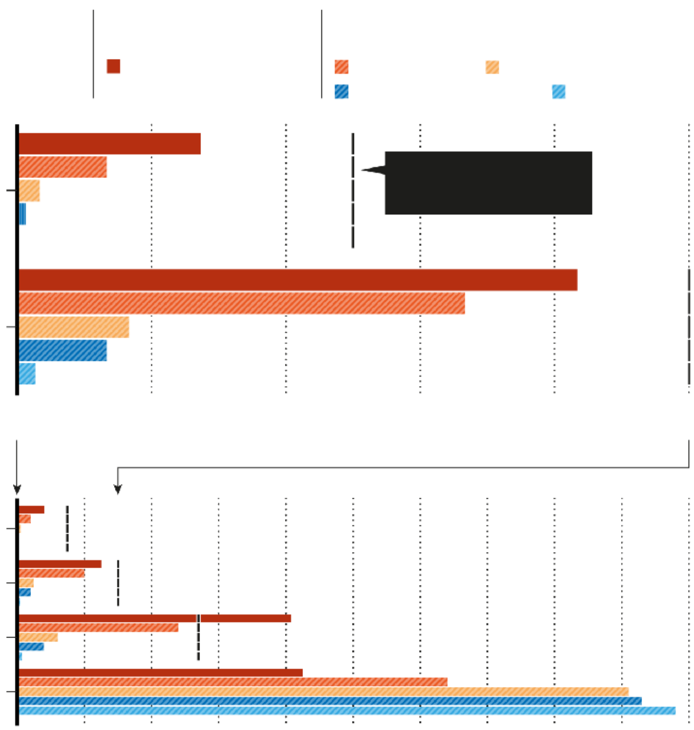
Un

Go

('r

Ad

Da



2

5

of to

Ad

Da

OE

Ge

)

)

)

)

0

Diversity at top Indian institutions

Universities in India are failing to meet government quotas for marginalized communities in academia. The government provides detailed national figures for undergraduate and master's students, but not for academic levels above those; Nature collated figures from right-to-information requests. These data show that the representation of marginalized castes and other groups falls at higher academic levels.

of total (%)

Nationwide

*Science, technology, engineering and mathematics; government data 2019–20. †Data obtained from Nature informa requests, 2020 (latest available at time of collection). PhD: IITs: Delhi, Bombay (Mumbai), Madras (Chennai), Kanp Kharaghur. All other career levels: IITs: Delhi, Bombay (Mumbai), Kanpur, Kharagpur. ‡Other Backward Classes. T government refers to 'Other Backward Classes' as educationally and socially marginalized groups that are not Sched Castes (Dalit) or Scheduled Tribes (Adivasi). §'General' is the government term for all groups that are not OBC, Da Adivasi: that is, historically privileged castes.

figures

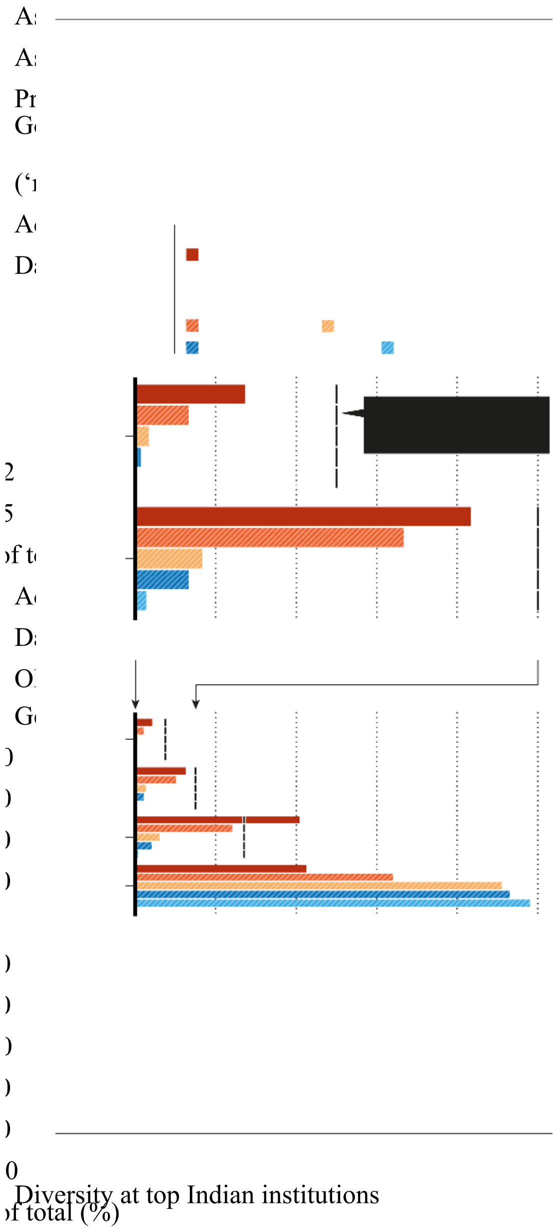
Top-ranked IITs

(Indian Institutes of Technology)†

Career level

Undergrad (STEM)*

PhD student



Diversity at top Indian institutions

Universities in India are failing to meet government quotas for marginalized communities in academia. The government provides detailed national figures for undergraduate and master's students, but not for academic levels above those. Nature collated figures from right-to-information requests. These data show that the representation of marginalized castes and other groups falls at higher academic levels.

Nationwide figures

Career level

Undergrad (STEM)*

Top-ranked IITs

(Indian Institutes of Technology)†

Career level

*Science, technology, engineering and mathematics; government data 2019–20. †Data obtained from Nature information requests, 2020 (latest available at time of collection). PhD: IITs: Delhi, Bombay (Mumbai), Madras (Chennai), Kanpur, Kharaghur. All other career levels: IITs: Delhi, Bombay (Mumbai), Kanpur, Kharagpur. ‡Other Backward Classes. The Indian government refers to 'Other Backward Classes' as educationally and socially marginalized groups that are not Scheduled Castes (Dalit) or Scheduled Tribes (Adivasi). §'General' is the government term for all groups that are not OBC, Dalit or Adivasi: that is, historically privileged castes.

PhD student

Assistant professor

Associate professor

Professor

Government quota

(‘reservation’ policy)

Adivasi

Dalit

2

5

of total (%)

Adivasi

Dalit

OBC‡

General§

“This is deliberate” on the part of institutes that “don’t want us to succeed”, says Ramesh Chandra, a Dalit, who retired as a senior professor at the University of Delhi last June. Researchers blame institute heads for not following the reservation policies, and the government for letting them off the hook.

Diversity gaps are common in science in many countries but they take different forms in each nation. The situation in India highlights how its caste system limits scientific opportunities for certain groups in a nation striving to become a global research leader.

0

of total (%)

India’s government publishes summary student data, but its figures for academic levels beyond this don’t allow analyses of scientists by caste and academic position, and most universities do not publish these data. In the past few years, however, journalists, student groups and researchers have been gathering diversity data using public-information laws, and arguing for change. *Nature* has used some of these figures, and its own information requests, to examine the diversity picture. Together, these data show that there are major gaps in diversity in Indian science institutions.

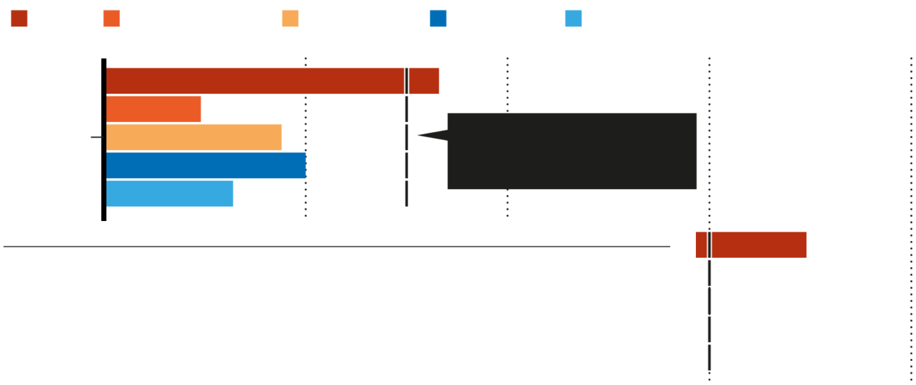
*Science, technology, engineering and mathematics; government data 2019–20. †Data obtained from Nature information requests, 2020 (latest available at time of collection). PhD: IITs: Delhi, Bombay (Mumbai), Madras (Chennai), Kanpur, Kharaghur. All other career levels: IITs: Delhi, Bombay (Mumbai), Kanpur, Kharagpur. ‡Other Backward Classes. The Indian government refers to ‘Other Backward Classes’ as educationally and socially marginalized groups that are not Scheduled Castes (Dalit) or Scheduled Tribes (Adivasi). §‘General’ is the government term for all groups that are not OBC, Dalit or Adivasi — that is, historically privileged castes.

Early barriers

The challenges start in schools and feed through to university admissions. Adivasis and Dalits are under-represented in undergraduate science courses [illegible] rates, higher-education survey data show (see ‘Undergraduate students in India’).

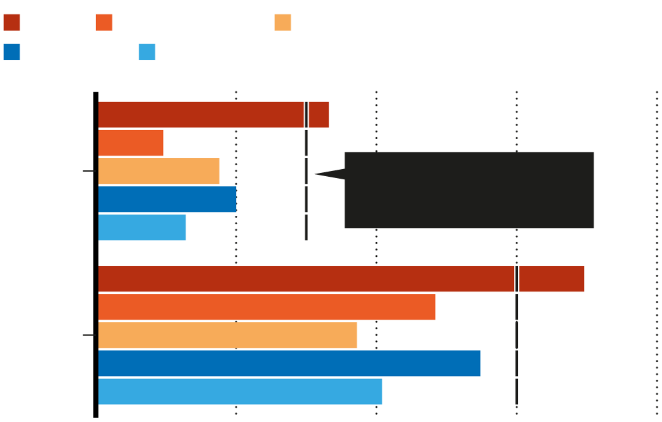
Diversity at top Indian institutions

Universities in India are failing to meet government quotas for marginalized communities in academia. The government provides detailed national figures for undergraduate and master's students, but not for academic levels above those; Nature collated figures from right-to-information requests. These data show that the representation of marginalized castes and other groups falls at higher academic levels.



Undergraduate students in India

At undergraduate level, Adivasis and Dalits are under-represented in science courses but not arts courses (2019–20 government data).



Government quota ('reservation' policy)

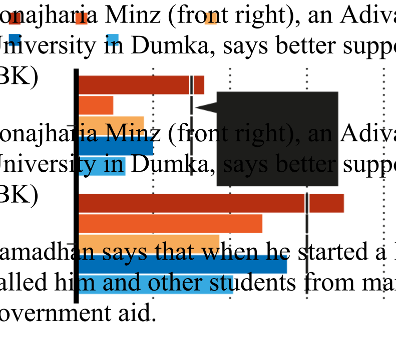
Dalit

Nationwide figures

Career level

That is not because arts courses are more popular, but because teachers and mentors specializing in science are rare in the rural high schools attended by these students, especially Adivasis, says Sonajharia Minz, a computer scientist and vice-chancellor of Sido Kanhu Murmu University in Dumka, eastern India. (Minz is the second Adivasi woman to hold a vice-chancellorship in India.)

Sonajharia Minz (front right), an Adivasi, computer scientist and vice-chancellor of Sido Kanhu Murmu University in Dumka, says better support systems are needed for marginalized groups. Credit: Balkishor Tudu (BK)



Sonajharia Minz (front right), an Adivasi, computer scientist and vice-chancellor of Sido Kanhu Murmu University in Dumka, says better support systems are needed for marginalized groups. Credit: Balkishor Tudu (BK)

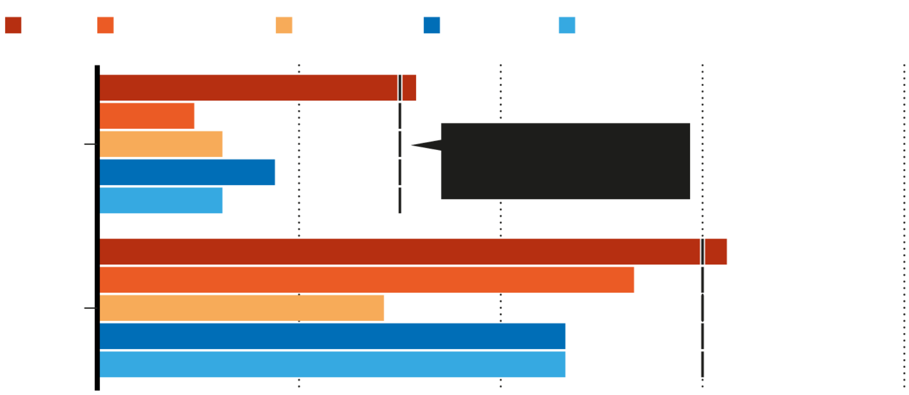
Samadhan says that when he started a life-science bachelor's degree in 2009, students from privileged castes often called him and other students from marginalized communities "free off" — a slur referring to students on government aid.

Another marginalized group, termed 'other backward classes' or OBCs, makes up around 44% of science students and 31% of medical students at the undergraduate level. India's national proportion of OBCs is not known, because the country's census does not count them. But a government household survey from 2006 suggests they make up around 41% of the population. (Reservation policies require academic institutions to have 27% of admissions or recruitments from OBCs.)

Undergraduate students in India

At undergraduate level, Adivasis and Dalits are under-represented in science courses but not arts courses (2019–20 government data).

In 2012, Samadhan progressed to a master's degree at a high-ranking university in western India. Student diversity at master's level is slightly lower than at the undergraduate level, data show (see 'Master's students in India').

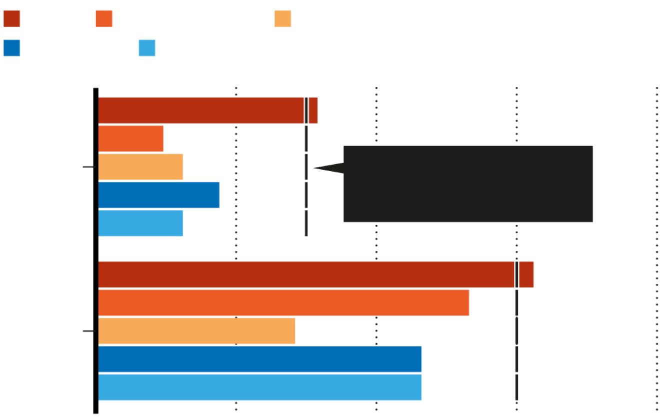


Master's students in India

At master's level, the under-representation of marginalized castes or groups in Indian universities worsens slightly, compared with undergraduates (2019–20 government data).

Subject

Medicine



Master’s students in India

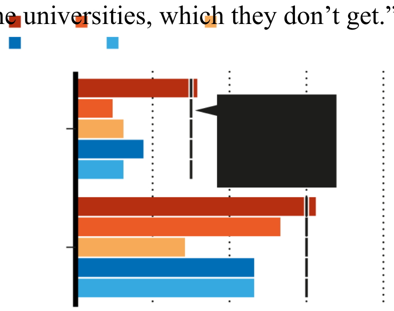
At master’s level, the under-representation of marginalized castes or groups in Indian universities worsens slightly, compared with undergraduates (2019–20 government data).

Subject

Medicine

Arts

During his master’s, Samadhan often considered dropping out because he felt intimidated by the English-speaking culture and intense coursework that he wasn’t used to. This is a common experience among students from underprivileged communities, says Akash Gautam, a Dalit and an assistant professor of neuroscience at the University of Hyderabad. “Many of them are first learners in their families. They need more time and support from the universities, which they don’t get.”



Master’s students in India

At master’s level, the under-representation of marginalized castes or groups in Indian universities worsens slightly, compared with

und
gov
Sub
Me
Art
Eng
g
Sci
Tec
y
Gov
ent
('re
ion
pol
Adi
Dal

0

5
0
n of total



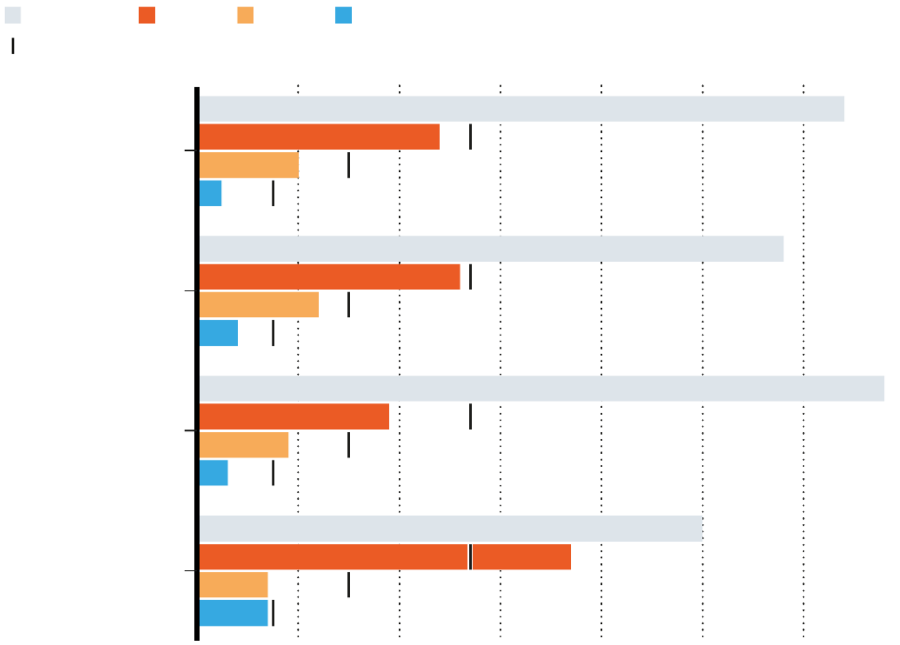
Akash Gautam, a Dalit and an assistant professor of neuroscience at the University of Hyderabad. Credit: Asiqul Ali

PhD dip

At the PhD level, the proportion of marginalized communities dips further — particularly at elite institutions. Data for PhD courses in 2020 at five high-ranked IITs, collected by *Nature*, show an average of 10% representation for Dalits and 2% for Adivasis — slightly lower than the average for five mid-ranking IITs. India's top-ranked university, the Indian Institute of Science (IISc) in Bengaluru, fares badly too (see 'PhD students in India').

Akash Gautam, a Dalit and an assistant professor of neuroscience at the University of Hyderabad. Credit: Asiqul Ali

Akash Gautam, a Dalit and an assistant professor of neuroscience at the University of Hyderabad. Credit: Asiqul Ali



PhD students in India

Data from a dozen Indian universities and institutes, sourced through information requests, show under-representation of marginalized castes in PhD programmes.

Dalit

Adivasi

General*

OBC†

Government quota (‘reservation’ policy)

Top−ranked IITs‡

Mid-ranked IITs§

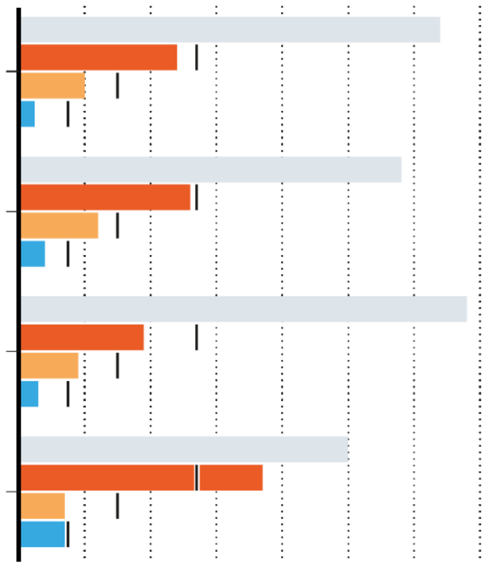
IISc||
Banaras Hindu

University¶

)

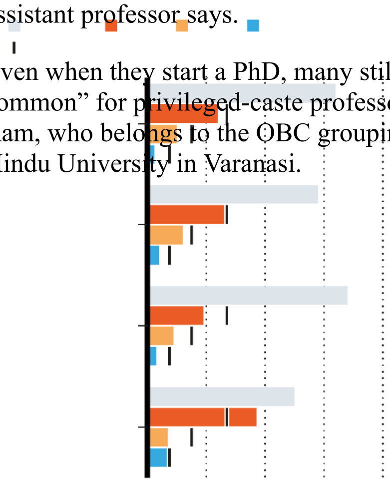
)

)
)
)
)
)
of t



*(’. ‡Delhi,
Bo d, Patna,
Gu an official
ra data shown.
A me of
co Egalitarians
gr

“Let’s face it, a PhD is somewhat of an elite pursuit” requiring financial support from families, says an IIT Delhi assistant professor from a privileged caste, speaking on condition of anonymity. Students from marginalized castes also often lack the recommendation networks and interview training to get recruited to PhD programmes, the assistant professor says.



Even when they start a PhD, many still struggle to find a good mentor willing to take them on. It is “quite common” for privileged-caste professors not to supervise students from marginalized communities, says Kirpa Ram, who belongs to the OBC grouping and is an assistant professor of environmental sciences at the Banaras Hindu University in Varanasi.

nd institutes, sourced
ler-representation of
s.



Kirpa Ram, who belongs to the OBC group, is an environmental scientist at the Banaras Hindu University in Varanasi. Credit: Kirpa Ram

An Adivasi PhD student from a science department at Delhi University, for instance, told *Nature* that in 2018, when she approached a privileged-caste professor to be her supervisor, he responded that she was a “quota candidate” and could find a supervisor anywhere.

Gautam and other scientists told *Nature* that sometimes privileged-caste professors don’t provide the extra mentoring that students from underprivileged backgrounds might need — leading some to quit before completing their PhDs. “It’s a tactic,” Gautam says.

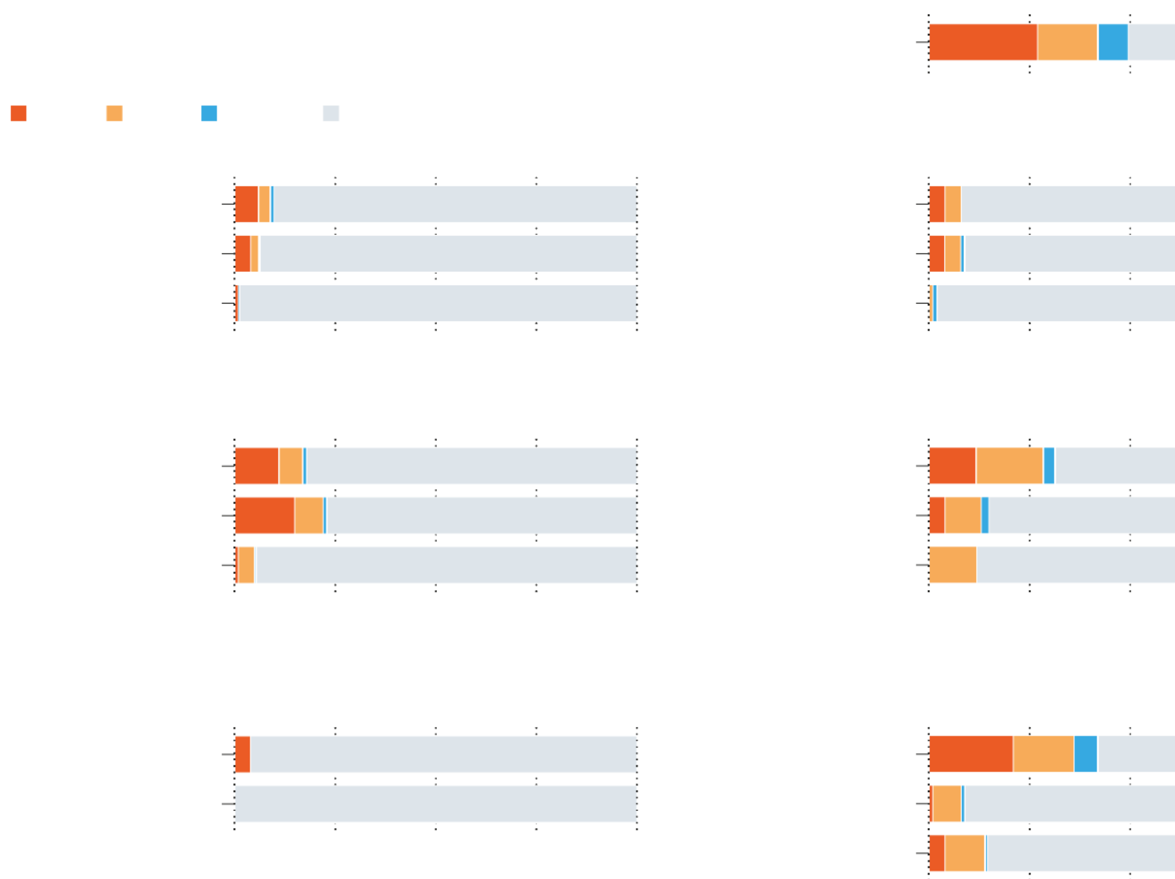
*Groups that are not Dalit, Advasi or OBC. †’Other backward classes’. ‡Delhi, Bombay (Mumbai), Madras (Chennai), Kanpur, Kharagpur. §Dhanbad, Patna, Guwahati, Ropar, and Goa. IIT Goa was created too recently to have an official ranking. ||Indian Institute of Science, Bengaluru. ¶Faculty of Science data shown. All data from Nature information requests, 2020 (latest available at time of collection), except for Patna, Goa and Guwahati from requests by the Egalitarians group.

Staffing fall

Very few doctoral students from marginalized castes reach staff positions in elite institutes (see ‘Faculty members in India’). At higher-tier IITs and the IISc, 98% of professors and more than 90% of assistant or associate professors are from privileged castes, *Nature* found. In the Mumbai-based Tata Institute of Fundamental Research (TIFR), all professors are from the privileged castes, the data it provided to *Nature* suggest. TIFR belongs to a class of federally funded ‘Institutions of Excellence’ exempt from following reservation policies.

Kirpa Ram, who belongs to the OBC group, is an environmental scientist at the Banaras Hindu University in Varanasi. Credit: Kirpa Ram

Kirpa Ram, who belongs to the OBC group, is an environmental scientist at the Banaras Hindu University in Varanasi. Credit: Kirpa Ram



Faculty members in India

Data from 13 elite institutions in India show that a very low proportion of faculty members are from marginalized castes, especially at the assistant professor level and above.
Government quota

(‘reservation’ policy)

All levels

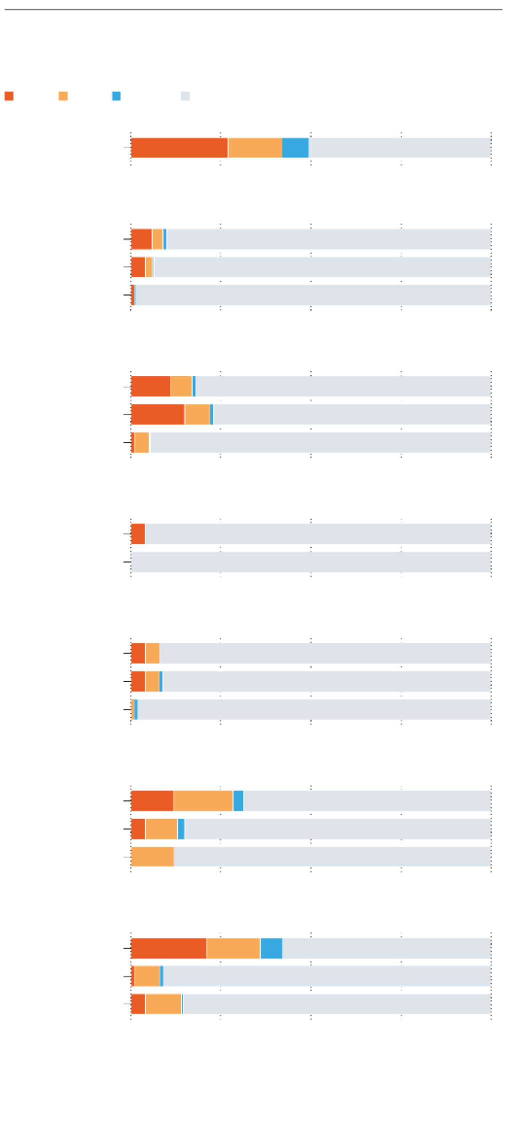
5

)

5

0

Dalit



Associate professor

Faculty members in India

Associate professor

Professor

Data from 13 elite institutions in India show that a very low proportion of faculty members are from marginalized castes, especially at the assistant professor level and above.

Professor

Dalit

Adivasi

OBC†

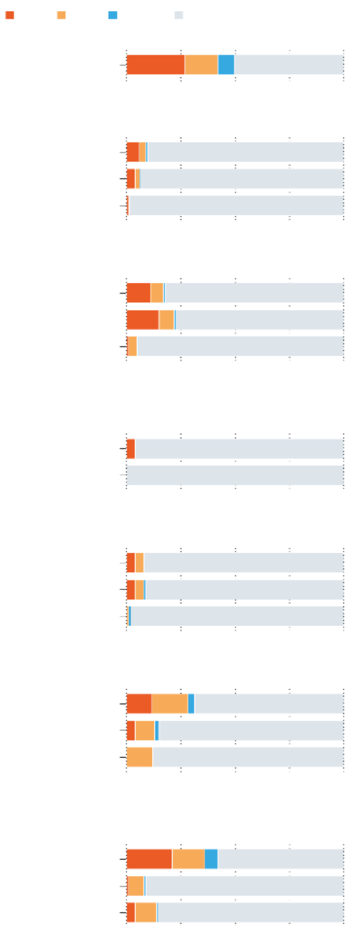
General*

Government quota (‘reservation’ policy)

All levels

Faculty members in India

Data from 13 elite institutions in India show that a very low proportion of faculty members are from marginalized castes, especially at the assistant professor level and above.



ch, Mumbai

OBC. †'Other backward classes'. ‡Delhi, Bombay (Mumbai), Kanpur, Kharag

ıd, Patna, Guwahati, Varanasi, and Goa. IIT Goa was created too recently to ha

Bengaluru. All data from Nature information requests, 2020 (latest available at

l Guwahati from requests by the Egalitarians group.

5 Jawaharlal Nehru University, New Delhi
0 Assistant professor
Associate professor
n ethnicities (%)
Top-ranked IITs‡
Professor
Assistant professor
5 Associate professor
0 Professor
5
0
n ethnicities (%)
5 Banaras Hindu University (All departments)
0 Assistant professor
Associate professor
n ethnicities (%)
Mid-ranked IITs§
Professor
Assistant professor
5 Associate professor
0 Professor
5
0
n ethnicities (%)
5
0
n ethnicities (%)
Tata Institute of Fundamental Research, Mumbai

*Groups that are not Dalit, Advasi or OBC. †'Other backward classes'. ‡Delhi, Bombay (Mumbai), Kanpur, Kharagpur. (IIT Madras didn't provide faculty data). §Dhanbad, Patna, Guwahati, Varanasi, and Goa. IIT Goa was created too recently to have an official ranking. ||Indian Institute of Science, Bengaluru. All data from Nature information requests, 2020 (latest available at time of collection), except for Patna, Goa and Guwahati from requests by the Egalitarians group.

Associate professor
Professor
5
0
5
0
n ethnicities (%)
IISc||
Assistant professor
Associate professor
Professor
5
0
5
0

n ethnicities (%)

Jawaharlal Nehru University, New Delhi

Assistant professor

Associate professor

Professor

5

)

5

0

n ethnicities (%)

Banaras Hindu University

(All departments)

Assistant professor

Associate professor

Professor

Some premier institutes are doing a little better. At the Council of Scientific and Industrial Research, a group of 38 labs, 18% of researchers (combining senior staff grades) were Dalits and 4% Adivasis, according to data *Nature* received from 31 of those labs.

5

)

5

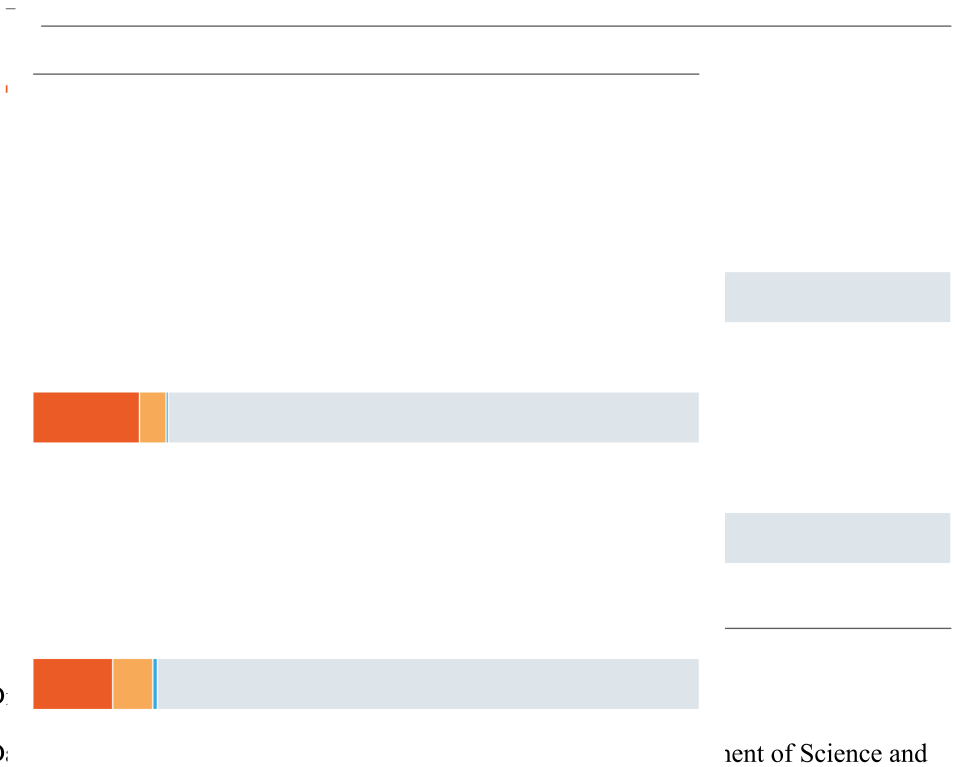
0

Funding mismatch

Most of India's research funders don't disclose data on funding by caste, or don't collect them in the first place. But the Department of Science and Technology (DST), one of India's two main science-funding government agencies, did share data with *Nature* on postdoctoral researchers whom it had awarded INSPIRE Faculty Fellowships — positions aimed at supporting young talent, which represent an important but small part of the DST's total funding. Between 2016 and 2020, 80% of recipients were from privileged castes, just 6% were Scheduled Castes (Dalits) and less than 1% Scheduled Tribes (Adivasis). The DST says the selection was "strictly based on merit". Funding given by the DST's Technology Development and Transfer Division showed a similar pattern (see 'Diversity in India's research funding'). In both cases, the DST didn't share application success rates.

n ethnicities (%)

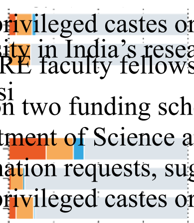
*Groups that are not Dalit, Advasi or OBC. †'Other backward classes'. ‡Delhi, Bombay (Mumbai), Kanpur, Kharagpur. (IIT Madras didn't provide faculty data). §Dhanbad, Patna, Guwahati, Varanasi, and Goa. IIT Goa was created too recently to have an official ranking. ||Indian Institute of Science, Bengaluru. All data from Nature information requests, 2020 (latest available at time of collection), except for Patna, Goa and Guwahati from requests by the Egalitarians group.



Data on two funding schemes, shared with Nature by India's Department of Science and Technology (DST) under information requests, suggest most money goes to researchers from privileged castes or groups.

Diversity in India's research funding

INSPIRE faculty fellowship for postdoctoral research, 2016–20



Data on two funding schemes, shared with Nature by India's Department of Science and Technology (DST) under information requests, suggest most money goes to researchers from privileged castes or groups.

INSPIRE faculty fellowship for postdoctoral research, 2016–20

Adivasi
Dalit
0.7%
OBC
12.0%
General
80.0%

Faculty members in India

DST Tech Development and Transfer Division grants 2016–20

Data from 13 elite institutions in India show that a very low proportion of faculty members are from marginalized castes, especially at the assistant professor level and above.

Adivasi
Dalit
OBC
16.0%
General
6.0%
Dalit
4.0%
Adivasi
OBC
General*
Government quota ('reservation' policy)

In the past few years, groups of activists, students and researchers have been pushing institutions not only to follow quotas but also to better support researchers from underprivileged communities. "We are doing this to hold a mirror to these institutes to show how ugly they are," says a spokesperson for Egalitarians, an organization that tries to collect and publish diversity data.

The issue is part of a wider recognition of how privileged-caste groups have been discriminating against marginalized communities — emboldened by India's pro-Hindu prime minister Narendra Modi, say some academics who spoke to *Nature* but did not want their names on record.

Some minority religious groups in India, such as Muslims — who are present across a variety of social and caste divisions, including Dalits — also face structural inequities in society. According to the 2011 census, Muslims make up around 14% of the country's population, but only 5.5% of those enrolled in higher education in 2019–20 were Muslims, survey data suggest.

Because no more detailed data were available, *Nature* asked IITs and other institutes for figures to do with Muslim representation. Most replied that they didn't have the figures, but the sparse data that a few institutions shared suggest that Muslims are under-represented in elite academic institutes. In 2020, Muslims made up less than 5% of PhD students in IIT Madras in Chennai and less than 1% of science-teaching faculty in IIT Kharagpur; both are prestigious institutes. However, in IIT Dhanbad, a mid-ranking institute in a region where Muslims are not unusually numerous, 55% of PhD students were Muslims.

The rising criticism of under-representation and discrimination in academic institutions, particularly around caste, is prodding some institutes into action. India's Ministry of Education, which didn't respond to *Nature*'s request for an interview, has several times since 2019 told federally funded institutes, including the IITs, to comply with reservation norms when recruiting teachers.

In 2019, Modi's government expanded the reservation quotas by 10% to cover lower-income people not part of marginalized castes or groups, who would otherwise fall in the 'General' category; they would be categorized under 'Economically Weaker Sections' of society. The extension is controversial but, after legal challenges, was upheld in a November 2022 ruling by India's Supreme Court.

Representatives at four IITs talked on record to *Nature* for this story; others didn't respond. "Equating under-representation with discrimination is incorrect. There is no discrimination," said Neela Nataraj, a mathematician and dean of faculty affairs at IIT Bombay in Mumbai. She accepted that the institute had a shortage of students and faculty members from some social categories, but said it was on a mission to improve representation through recruitment without compromising on quality, and through encouraging more students from under-represented communities to start PhDs. Angelie Multani, a professor of literature who was appointed in August 2022 as IIT Delhi's first dean to increase diversity on campus, says the institute recognized that, like others, it had "under-representation of marginalized sections of society" and was working to improve the situation through measures such as hiring drives. And Amalendu Chandra, dean of faculty affairs at IIT Kanpur, says that the institute had offered appointments to 48 teachers from marginalized communities in the past year. (The institute has 413 faculty members, according to its website).

At IIT Goa, Amaldev Manuel, a computer scientist and chair of PhD admissions in 2022, noted that the institution's acceptance rate for PhD applicants from marginalized communities was higher (at above 1%) than for the 'general' category (below 1%), even though it received fewer applications from people of less privileged castes.

Some researchers, such as Ramesh Chandra, doubt that diversity initiatives by institutions will make a big dent until India's government takes action against institutes for violating reservation policies. "You have to take punitive action against the [institutes'] directors," says Chandra. "Remove them."

At the very least, says Ram, the government should require universities to publicly disclose diversity data and monitor compliance. And Minz thinks that for the situation to change, support systems need to be created at every step from school education to high-level recruitments in academia — such as training on grant-proposal-writing and communication skills for researchers recruited from marginalized communities. "The playing field is not equal at any stage," she says.

For Samadhan, the marker of change would be more personal. "The day I would be able to say my full name without hesitation in an institute, I will feel that equality has arrived," he says.

This article is also available as a pdf version.

Diversity in India's research funding

Data on two funding schemes, shared with Nature by India's Department of Science and Technology (DST) under information requests, suggest most money goes to researchers from privileged castes or groups.

INSPIRE faculty fellowship for postdoctoral research, 2016–20

Adivasi 0.7%
OBC 12.0%
General castes 80.0%
Dalit 6.0%

DST Tech Development and Transfer Division grants 2016–20

OBC 16.0%
Adivasi 0.4%
General 81.0%
Dalit

Mid-ranked
Assistant professor
Associate professor
Professor
Not known

Tata Institute of Fundamental Research, Mumbai

SPRINGER NATURE

Springer Nature

A 0% Associate
- Privacy Policy
- Use of cookies
professor
- Legal notice
Professor
- Terms & Conditions
- Accessibility statement

© 2023 Springer Nature Limited. All rights reserved.

5

Top Built with **Shorthand**

0

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00015-2

5

| Section menu | Main menu |

)0

known
)

IISc||
Assistant

professor
Associate

professor

Professor

5

0

5

)0

known
)

Jawaharlal Nehru
University,

New Delhi
Assistant

professor
Associate

professor

Professor

5

0

5

)0

known
)

- NEWS FEATURE
- 10 January 2023

Colonoscopies save lives. Why did a trial suggest they might not?

A major clinical study raised questions about one of the most celebrated cancer-screening procedures available, but a close look at the data tells a different story.

- Emily Sohn

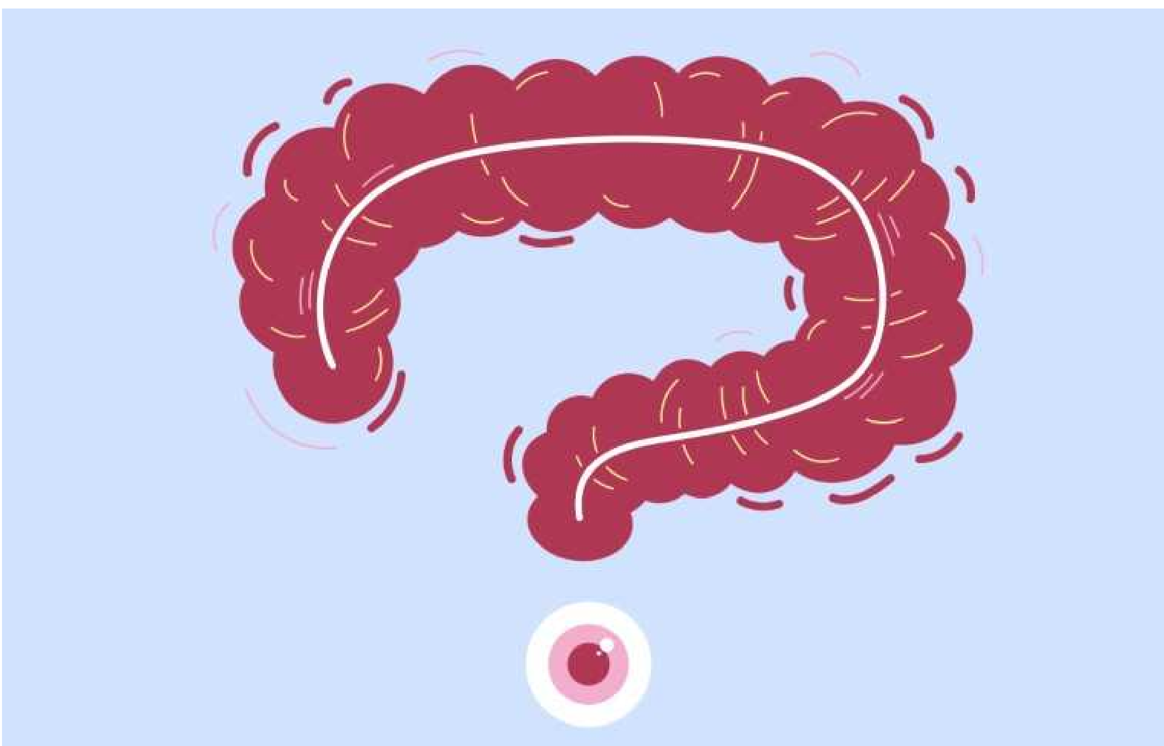


Illustration by Jasiek Krzysztofiak

It was an uncomfortable moment for people who perform colonoscopies. In October, a massive randomized clinical trial in Europe presented its initial

results[1], which suggested that, as a screening tool, colonoscopies don't save as many lives as expected.

Researchers were perplexed because the procedure had long been considered a true success story in cancer screening. But after the study results were compared with data from other trials, colonoscopy to be seemed less effective than simpler screening methods that assess only part of the colon. Jason Dominitz, a physician at the Veterans Health Administration in Seattle, Washington, says it was like suggesting that mammography on only one breast was better than scanning both. It didn't make sense.

A media frenzy followed, and headlines were blunt, declaring that colonoscopies might not be effective or prevent deaths at all. But when Dominitz dug deeper, the trial results reflected where and how the study was conducted and the complexity of the questions it was trying to answer. "It is really important to not just read the headline," says Dominitz, who is also director of the colorectal-cancer screening programme run by the US Department of Veteran Affairs.



The controversial embryo tests that promise a better baby

A closer look at the European study, on its own and in the context of other studies, shows that colonoscopies do in fact substantially reduce the risk of developing colorectal cancer and dying from it. They are still considered by

many experts to be one of the best ways to screen for the disease. But for any screening procedure, there are trade-offs both for individuals and at the public-health level. As scientists are working out the details of which tests to recommend, the reaction to the study illustrates how difficult it is to interpret and communicate research on cancer screening.

“It’s really important to look at all the evidence in totality,” says Jennifer Croswell, a public-health researcher at the National Cancer Institute (NCI) in Bethesda, Maryland, who specializes in cancer screening. “This was a complicated trial to sort through.”

Benefits of catching cancer early

Colorectal cancer is the third most common form of cancer and the second leading cause of cancer deaths globally, behind lung cancer. In 2020, 1.9 million people were diagnosed with it and 900,000 died from it around the world. Rates are highest in high-income countries, but are growing in low- and middle-income countries.

There is plenty of evidence to show that detecting the disease early saves lives. If colorectal cancer is found before it can spread, the 5-year survival rate is about 91%, according to the NCI, compared with 15% if the cancer has metastasized. There are signs that screening programmes have made a difference, particularly in the United States. In 2000, 38% of US adults over 50 were getting screened; that figure rose to 66% by 2018. Colorectal cancer rates dropped in that period and deaths fell from about 20 to 13 per 100,000 people.

In 1995, the US Preventive Services Task Force (USPSTF) began to recommend that colorectal-cancer screening start at age 50. But the vast increase in screening is often credited in part to the television news presenter Katie Couric, who broadcast her own colonoscopy in 2000 after her husband died from the disease. Other celebrities have since promoted the procedure, and the USPSTF lowered the recommended starting age for screening to 45 in 2021.

Colonoscopy has long been the most popular form of colorectal-cancer screening in the United States. It is highly valued because it allows doctors to examine the entire colon for signs of cancer and to remove polyps — abnormal growths that can become cancerous — during the process. But scientists have struggled to compare colonoscopy's effectiveness with that of less invasive methods, such as external imaging, stool-sampling techniques or flexible sigmoidoscopy, which looks at only half the colon.

Colonoscopy is much less common in Europe, in part owing to questions about whether the test is too invasive and expensive to be worth recommending, says Michael Bretthauer, a gastroenterologist at the University of Oslo. To address these questions, he and his colleagues planned a randomized trial of colonoscopies. Starting in 2009, they recruited more than 84,000 people aged 55 to 64 from Norway, Poland and Sweden. Some were invited to get screened. Others received their usual health care but no such invitation.



Colonoscopy as a screening tool for colorectal cancer is not as common in Europe as it is in the United States.Credit: Jean-Paul Pelissier/Reuters via

Alamy

With about ten years of follow-up data, Bretthauer and colleagues released their attention-grabbing results in October 2022, seemingly suggesting that colonoscopies had a smaller benefit than expected. There was just an 18% reduction in the risk of developing cancer among those who had been invited to get colonoscopies, and no significant reduction in the risk of death.

But the study itself offered layers of interpretation that cast colonoscopies in a more favourable light. Overall, only 42% of people in the group that had been invited to get colonoscopies actually got one. If the compliance rate had been 100%, the researchers' analysis showed, the test would have reduced cancer risk by 31% — from 1.22% to 0.84% — and it would have reduced the risk of death from colorectal cancer by 50% — from 0.3% to 0.15%.

Those benefits are significant, says Chyke Doubeni, a family doctor and colonoscopy researcher at the Ohio State University in Columbus, and there are reasons to think that they could be larger in other circumstances, especially in populations that experience disproportionately high rates of the disease. And despite the huge scale of the European study, ten years of follow-up is a relatively short period of time for colorectal-cancer development, says Amy Knudsen, who studies disease simulation models to inform cancer-care policy at Massachusetts General Hospital and Harvard Medical School, both in Boston. "I think we're only going to see the impact of colonoscopy increase the longer we follow up," she says. The European study is continuing to track participants.

Dominitz also notes that roughly one-third of endoscopists in the study found polyps in less than 25% of colonoscopies performed. Typically, more than half of people develop polyps in their lifetimes, he says. The low rate of detection, primarily in Sweden, reflects relatively low risks of colorectal cancer in that country for reasons that have yet to be clarified, Bretthauer says.



Could CAR-T-cell therapy offer hope to children with cancer?

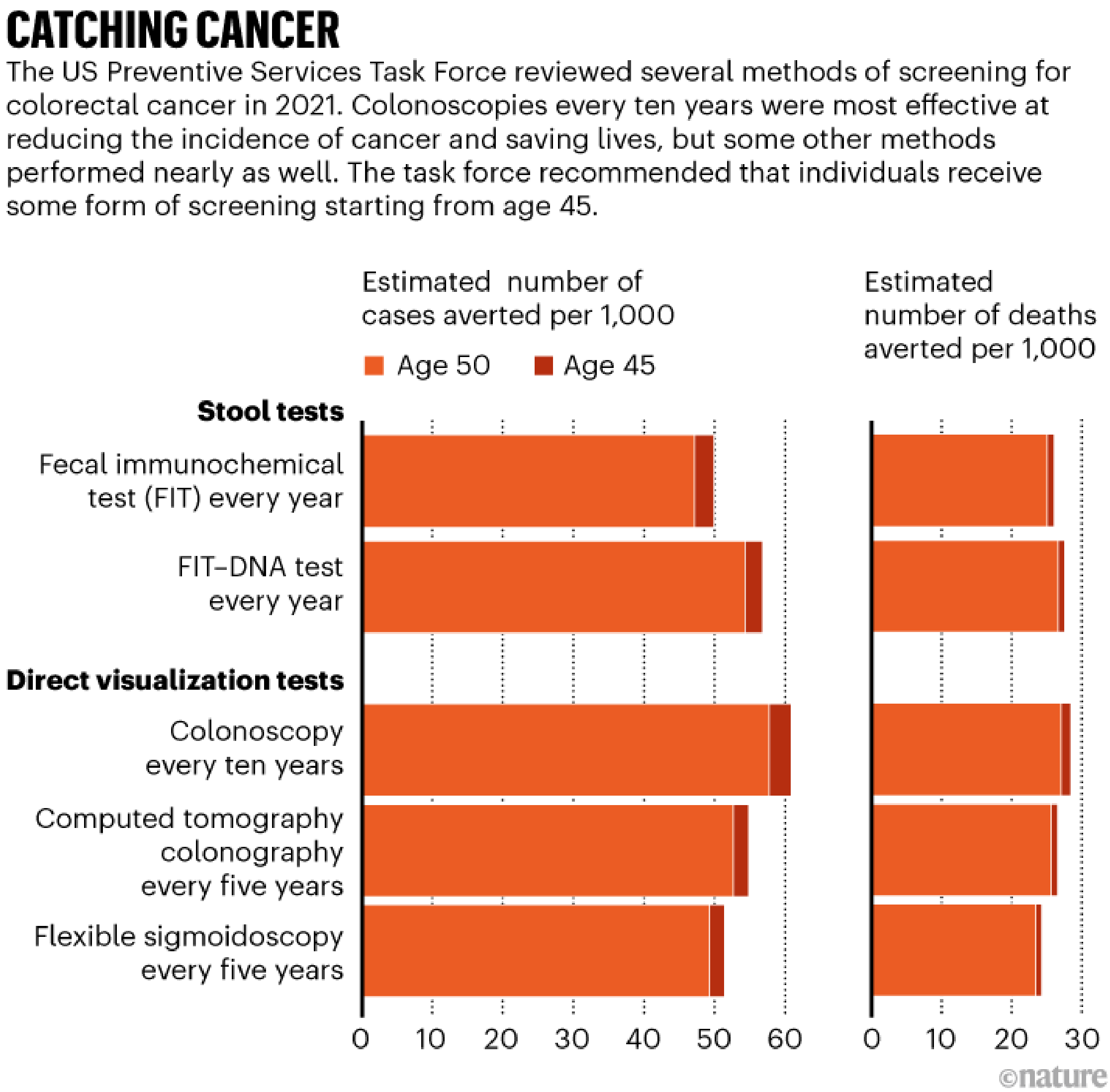
Some of the confusion about the study's results might be related to the kinds of question that it was trying to answer, which came from both a public-health and an individual-health perspective, says Bretthauer. "If I talk to you as a patient, I would say, 'You could probably reduce your risk of getting colon cancer by 31%,'" he says. "If I talk to a politician who says, 'I'm wondering if I should start to introduce this,' I would say, 'Well, you can expect an 18% effect in your city.'" But that distinction was not often picked up in coverage of the study, says Knudsen.

"In the US media, I feel like the headlines were all talking about the trial results showing that colonoscopy was not good, and I don't think that's what the trial shows at all," she says. "I was very disappointed."

Other evidence

No single study can be used in isolation to evaluate something as complicated as cancer screening. For years, evidence has accumulated in support of colonoscopies, including several large studies that have suggested that the procedure has major benefits in some cases. In a 2021 review[2], the USPSTF found that, across observational and modelling studies, the risk of developing colorectal cancer is lower for people who get screening

colonoscopies, with estimates ranging from a 40% to 69% reduction, and that their risk of death is also lower, with estimates ranging from a 29% to 88% reduction (see ‘Catching cancer’). For every 1,000 people screened with colonoscopy, modelling suggests that as many as 28 lives are saved by early detection.



Source: US Preventive Services Task Force

Although the European study has been portrayed as disappointing, Doubeni says the results are consistent with previous data. Randomized trials are important, he adds, but observational studies have the benefit of reflecting real-world conditions. “I think there’s fairly good evidence at this point,

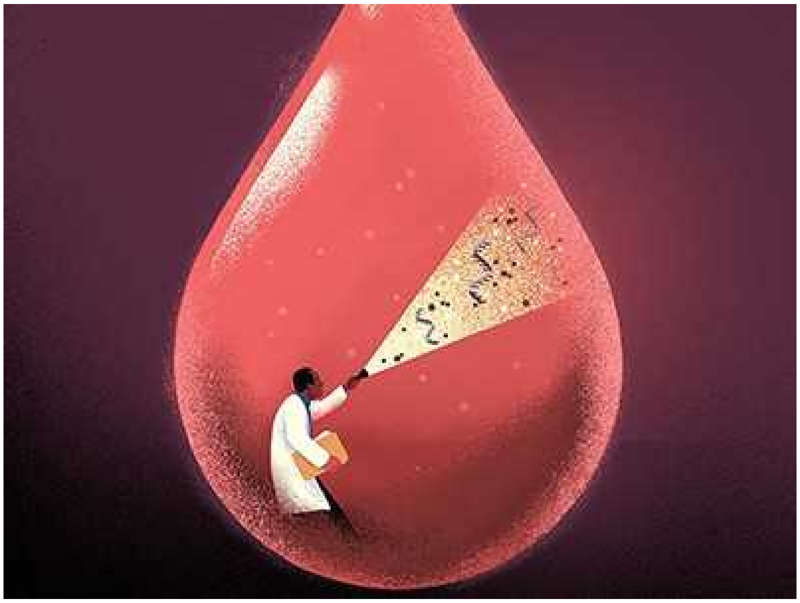
even outside of clinical trials, that colonoscopy is effective," he says. "I have no doubt about that."

There are other options for screening: CT colonography (which uses computed tomography imaging instead of a scope), flexible sigmoidoscopy and two main types of stool test. One, called a fecal immunochemical test (FIT), uses antibodies to detect signs of haemoglobin (a protein in red blood cells) in stool, which can be indicative of cancer. The other, FIT–DNA, detects haemoglobin and DNA that gets shed from the lining of the colon and rectum.

Each type of test comes with trade-offs between potential benefits, risks, discomfort, cost and logistics. Getting a colonoscopy, for example, requires going 24 hours without solid food and undergoing sedation while a skilled doctor inserts a flexible tube equipped with a camera into the rectum. Complications can happen — studies show that, for every 10,000 procedures, there are an average of 3 bowel perforations[2], which sometimes require another surgery to repair. And even without complications, recovery is not instantaneous.

Stool tests don't involve bowel prep or sedation, but they need to be done as often as once a year; by contrast, colonoscopies are needed once every 10 years as long as the results are negative. Any positive result from another type of test requires a colonoscopy to confirm it, Doubeni says. And false positives are common, with rates as high as 13% for FIT-DNA and up to 5% for FIT.

A study published last year[3] found that people who didn't get a follow-up colonoscopy after a positive FIT stool test were twice as likely to die as those who did. "There is no test to do for colon-cancer screening that will not require colonoscopy to be done for you to get a benefit," he says. "It is important to underscore this."



How to improve the diagnosis of prostate cancer

Modelling suggests that the benefits are similar across screening types. In an analysis for the latest USPSTF report by the NCI-funded Cancer Intervention and Surveillance Modelling Network[4], the number of lives saved ranged from 24 per 1,000 people screened with flexible sigmoidoscopy to 28 per 1,000 for colonoscopy. "They're all in the same ballpark," Dominitz says. "The maximum spread here is four lives, and for those four people and their families, of course this is huge. But they're all really close. It's not like one is really dominant over the other."

Trials that have tested screening methods head-to-head suggest that compliance has a lot to do with these comparable success rates. Colonoscopies might be the best at seeing everything in the colon, but fewer people are motivated to get them. In a Swedish trial called SCREESCO[5], 35% of people who were advised to get a colonoscopy got one, compared with 55% of people in the FIT group. And in Spain, a study[6] called COLONPREV randomized more than 53,000 people into two groups — one received advice to get a colonoscopy, and the other was told to undergo FIT stool screening every two years. After 10 years, 30 cancers had been detected in the colonoscopy group, compared with 33 in the FIT group — largely because of people's behaviour. Only 25% of people in the colonoscopy group got screened, compared with 34% in the FIT group.

Dominitz is co-leading a study that is directly comparing cancer outcomes in 50,000 veterans who will be screened with either an annual FIT test or colonoscopy. So far, adherence in the study has been high, and screening colonoscopies have found precancerous polyps in 46% of participants. It's the first study of its kind in the United States, and, Dominitz says, it should more clearly help people to make decisions about how to get screened. "It gets more to the question of the patient in front of you who says, 'I want to be screened. What should I do?'" Results will be available in 2028.

People looking for guidance now will get different advice depending on where they live. Canada and the United Kingdom recommend stool-based screening (starting at age 50 in Canada and 60 in the United Kingdom), whereas the push for colonoscopies is mostly in the United States.

Bretthauer thinks that people should make their own decisions with all of the information available; he chose to get a colonoscopy, but he has colleagues in other countries who did not, on the basis of the same data.

Offering choices seems to be key to increasing screening rates in people of all ages. In one study[7], 69% of people who were advised to get a colonoscopy or stool testing got screened, compared with 38% who were advised to get just a colonoscopy. "For a screening test to be successful, it has to be acceptable to patients," Croswell says, and colonoscopies can be hard to stomach. "There's a yuck factor to having a scope stuck up your rectum. People are not necessarily excited about that." Compliance could also suffer if the test is too costly or difficult to schedule.

Researchers are continuing to monitor evidence to refine guidelines — but specialists say that, for colorectal cancer, getting any kind of screening is better than nothing. It's still not clear which test is best, Dominitz says. But stool tests, sigmoidoscopy and colonoscopy are all effective, as long as people do them. "Screening only works," he says, "if it gets done."

Nature **613**, 235-237 (2023)

doi: https://doi.org/10.1038/d41586-023-00020-5

References

1. Bretthauer, M. *N. Engl. J. Med.* **387**, 1547–1556 (2022).

2. Lin, J. S., Perdue, L. A., Henrikson, N. B., Bean, S. I. & Blasi P. R. *JAMA* **325**, 1978–1998 (2021).

3. Zorzi, M. *et al. Gut* **71**, 561–567 (2022).

4. Knudsen, A. B. *et al. JAMA* **325**, 1998–2011 (2021).

5. Forsberg, A. *et al. Lancet Gastroenerol. Hepatol.* **7**, 513–521 (2022).

6. Quintero, E. *et al. N. Engl. J. Med.* **366**, 697–706 (2012).

7. Inadomi, J. M. *et al. JAMA Intern. Med.* **172**, 575–582 (2012).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00020-5

Opinion

- **Indigenous knowledge is key to sustainable food systems**
  [10 Jan 2023]
  • Agricultural sciences have for too long ignored traditional and local knowledge about crop plants and how best to grow them. That must change if the world is to ensure future food security.
- **Use WHO antibody standards for emergency vaccine testing** [10 Jan 2023]
  •

- COMMENT
- 10 January 2023

Indigenous knowledge is key to sustainable food systems

Agricultural sciences have for too long ignored traditional and local knowledge about crop plants and how best to grow them. That must change if the world is to ensure future food security.

- Alexandre Antonelli



Farmer Angelina Monday works on her plot of land in Uganda, where she grows beans and vegetables for her family. Credit: Mads Nissen/Politiken/Panos Pictures

I grew up in Campinas, a city in southeast Brazil. The apples there, cultivated from European varieties since the 1960s, tasted sweet. But, given the choice, I would always pick papayas grown in our garden. My father, who knew that growing a temperate fruit tree in a tropical country seldom worked, instead filled our garden with tropical ones, including two varieties of papaya. Meanwhile, drawing on knowledge from her Indigenous roots, my mother grew all sorts of herbs in pots around the house, which she used to treat ailments such as diarrhoea and indigestion.

Indigenous peoples and other local communities, who might have lived in a region for thousands or hundreds of years, respectively, have long acted as foragers, growers and shapers of nature[1]. In many parts of the world, the food production systems developed by such communities — from irrigated crops to agroforestry systems — have been the dominant food systems supporting regional economies, and feeding rural and urban areas alike[2].

For the past three decades, various efforts involving academic and industrial partners have explored how biodiversity in low- and middle-income countries could be exploited commercially — bioprospected — for new pharmaceuticals and crop varieties, and how benefits could be shared equitably. Yet there are huge power imbalances between the wealthy countries and large corporations seeking the products, and the biodiversity-rich but economically and technologically deprived countries and communities providing them. In practice, the benefits rarely reach the people who are the knowledge holders and guardians of biodiversity and agrobiodiversity[3].

Today, food production is the biggest driver of biodiversity loss and contributes heavily to climate change and pollution — the three components of the 'triple planetary crisis' recognized by the United Nations as requiring resolution if humanity is to create a viable future on this planet. As such, there has never been more need to establish how Indigenous and local knowledge can contribute to the building of resilient, sustainable and nutritious food systems in a way that is equitable.

As others have noted[4], people from Indigenous and local communities who provide their knowledge to research must be involved from the start, take the lead on projects whenever possible and receive tangible, long-lasting

benefits from them. But my work as director of science at the Royal Botanic Gardens, Kew, in London has convinced me that these goals are not enough. (Kew is collaborating with some 400 organizations in more than 100 countries to develop nature-based solutions to food insecurity, biodiversity loss and the effects of climate change.)

Fundamentally, there needs to be a transformation in the way agricultural science — indeed, all of science — is conducted. Assumptions about what counts as legitimate scientific knowledge must be questioned[5]. A greater appreciation of the wealth of information held as a result of humans living with and using species over hundreds or thousands of years must be developed. And the diverse needs of countries and communities across the world must be much better understood.

The move to monoculture

The ways in which plants are used by societies around the world have narrowed drastically over the past 500 years[6]. This has happened largely because of efforts to maximize yield and profit, pursued with little regard for Indigenous peoples and smallholder producers and the knowledge they hold.

Much of this narrowing occurred during clashes between Indigenous populations and imperial powers. In the Americas alone, around 90% of the continent's Indigenous population — some 56 million people — died as a result of conflict and disease between 1492 and 1600[7]. In subsequent centuries, such powers enslaved or displaced surviving communities by clearing swathes of tropical ecosystems to grow sugar cane, cotton, coffee, maize (corn) and other commodities[8] that they had 'discovered' in their colonies[9]. In their hunt for the highest yields of cotton, the sweetest maize and so on, colonists ignored many of the agricultural practices that had been developed for a rich collection of locally adapted crops[6].



Soya-bean harvesting in Brazil’s Cerrado, much of which was biodiverse wooded grassland.Credit: Andressa Anholete/Bloomberg via Getty

In the past century, Indigenous knowledge has been dismissed in different ways. Take the Green Revolution, a vast increase in the production of food grains in the decades after the Second World War. This resulted mainly from the introduction of high-yielding crop varieties and changes to agricultural practices, such as the use of machinery and chemical inputs.

By tripling cereal production within four decades, the spread of agricultural technologies helped to alleviate hunger and poverty in some places. However, in others it created food insecurity and exacerbated pollution, deforestation and the displacement of Indigenous and small-scale production systems.

Consider the Cerrado in South America, where some of my Indigenous ancestors lived sustainably for millennia — now home to around 100,000 Indigenous people who represent more than 80 ethnicities. Since the 1970s, more than 40% of the Cerrado’s naturally acidic soil has been transformed into agricultural land, in part through the addition of five tonnes of

pulverized chalk or limestone per hectare, among other inputs. About 20% of that land is used to grow soya beans to provide fodder for cattle farms that are mainly overseas[10]. Because of the heavy reliance on machinery, relatively few local people are employed in the fields, and leakage of pesticides into drinking water has been linked to deaths among farmers[11].

Between the 1960s and 1980s, only three energy-rich crops — first wheat, then rice and maize — came to dominate international trade, replacing regional species and varieties of pulses, cereals, fruits and vegetables. (Smallholder farmers continued to grow and supply the latter to rural and urban areas[6,8].)



How to include Indigenous researchers and their knowledge

Crops such as cassava, sorghum and certain millets were eventually added to breeding programmes in tropical regions, and from the 1980s, varieties developed under a global partnership called the Consultative Group for International Agricultural Research (CGIAR) were made available to national research centres. This meant that breeders could modify them so they were better adapted to local conditions[12]. Even so, Indigenous peoples often point out that, unlike the food systems created by the Green Revolution, their systems have always been diverse, resilient, sustainable,

nutritious and circular — meaning that resources are used efficiently with little waste and moderate levels of consumption[7,13].

As critics have noted, the Green Revolution was a top-down, state-led process, principally shaped by US priorities and ideologies[14] that downplayed the importance of Indigenous and local knowledge. Indeed, in his Nobel Peace Prize lecture in 1970, US agronomist Norman Borlaug — who spearheaded the Green Revolution — talked about solutions "to cure all ills of a stagnant, traditional agriculture".

Old solutions for new problems

Crop growers worldwide face enormous challenges. Plants are becoming less resistant to pests and diseases. Crops are more likely to be damaged or fail because of droughts, heatwaves, floods, salinization and rising sea levels. Soils are degraded; rivers and watersheds are polluted; and the diversity and abundance of crop pollinators are in steep decline. Meanwhile, food production contributes to around 37% of greenhouse-gas emissions[15]. And changes in land use associated with it are the greatest threat to terrestrial biodiversity[16].

For humanity to progress towards a sustainable world with a secure food supply, the data now clearly show that we must change diets, reduce waste, diversify food systems — in existing or reduced agricultural land — and develop more-circular ways to produce food[16,17]. It is also becoming apparent that 'climate-smart', environmentally sustainable food sources can come from underutilized plants and wild relatives of crop plants. Efforts to locate and manipulate such food sources should be guided by the communities who have long been using them, and who might be the main beneficiaries[18]. (Most current work on crops largely entails mapping and editing genes and their functions in only a few dozen species.)

Bananas are one example. The seemingly incurable fungal infection Panama disease is affecting dense banana plantations across the tropics, posing a major threat to the global trade of bananas, which is currently dominated by the Cavendish variety. Yet the disease does not affect crops in Africa and southeast Asia, where Indigenous and local communities grow hundreds of

varieties that are naturally resistant to it. In Uganda alone, nearly 100 varieties of banana contribute to the sustainable livelihoods of millions of people[19].



Enset trees surround buildings in southwest Ethiopia. The starch-filled stems are used in cooking, and the leaves as fodder and for roofing.Credit: Paul Wilkin RBG KEW

Likewise, the plant enset, also called the tree against hunger (*Ensete ventricosum*), is hardly known outside southwestern Ethiopia. But in this region, where it has probably grown for millennia, its starch-filled stems provide more than 20 million people with their main source of calories and nutrients. Its leaves are used to feed cattle, to provide shade and to build roofs; the stems provide fibres for basket-making; and when planted in rows, enset is used in place of fences to define land ownership and reduce conflicts. Perhaps most importantly in relation to climate change, enset is remarkably tolerant of drought and short-term temperature variations[20].

Almost 500 million Indigenous people, speaking at least 4,000 languages, occupy more than 25% of the global land surface. They often have the best ecological knowledge about the region they live in, and know which species

are most important for their communities[21]. They also hold taxonomic knowledge that has been neglected by scientists from elsewhere.

For instance, Indigenous Iban and Dusun communities in southeast Asia have long recognized that two similar-looking fruits, lumok and pingan, originate from two distinct plants. Yet for nearly two centuries, Western botanists had misclassified them as coming from a single tree species[22].

In principle, Indigenous and local knowledge could help to create training data sets. These could enable researchers and crop developers to find food sources that contain certain nutrients, that will tolerate anticipated climate shifts or that harbour resistance to emerging pests and pathogens. Phylogenomic models and artificial intelligence could then mine these data to predict the occurrence and function of genes that underlie useful traits across the tree of life[23]. And with genomic sequencing and gene-editing techniques becoming more accessible and affordable, more of the work to modify and cultivate species and varieties could happen at local and regional levels.

Routes to agrobiodiversity

Some might question why Indigenous and other communities would want to get involved in projects focused on the development of food crops after centuries of being exploited, mistreated and ignored.

Numerous efforts over the past three decades have been problematic for various reasons. Promises to share benefits equitably have not been kept owing to unequal power relations, unclear legislation around access to biological and genetic resources and the sharing of benefits, and complex financial mechanisms for transferring them[24].

Harms are ongoing. In 2007, a Dutch company gained patent rights for storing and processing a grain called teff, which has been used by Ethiopian people for thousands of years. Although an agreement for access and benefit sharing was signed in 2005, the expectations of multiple benefits to Ethiopian knowledge providers were not met. Only in 2019 did a court in the Netherlands rule that the patent was invalid.

I am under no illusion about what it will take to achieve true collaboration at scale — both at the individual and systemic level. Yet in my interactions with Indigenous people and local communities, people's generosity and willingness to work collaboratively has impressed me again and again.



Coffee berries from *Coffea stenophylla*, a species that communities in Sierra Leone helped researchers at the Royal Botanic Gardens, Kew, and their partners to find in the wild. It can tolerate higher temperatures than the commercial species *Coffea arabica*, which is under threat from climate warming.Credit: CRB Coffea, IRD-CIRAD

Although Indigenous communities have yet to receive the recognition they deserve, food-security workshops involving Indigenous people and local communities are beginning to shape research agendas[25] — such as those organized under the Colombia Bio programme involving multiple Colombian and British stakeholders, and co-led by colleagues at Kew. By adopting more-holistic and more-sustainable views on the management of

nature, and focusing on the relationships between human and non-human species[4], Indigenous peoples are challenging the status quo of economic growth at the cost of natural capital. Also, some Indigenous and local communities are already active in international debates about food security, biodiversity and climate change. Many are authoring key reports and voicing their perspectives, such as through the United Nations Food and Agriculture Organization[2,13] and during the UN biodiversity conference (COP15) held last month in Montreal, Canada.

With the backing of multilateral initiatives, such as the World Food Programme, and regional organizations, such as the Pan-African Council, funding agencies could promote collaboration of Indigenous and other local communities with crop-development scientists from Europe, the United States, China and other countries. The US National Science Foundation or the Horizon Europe programme, for example, could make funding calls specifically for biocultural research related to food security and non-conventional food crops. And grants could be evaluated partly on the basis of how much applicants propose to work on projects involving partnerships with Indigenous peoples and local communities, or on whether grants are initiated and led by collaborators from the global south.

Governments in high-income countries could earmark part of their official development-assistance funds or other sources to support knowledge exchange within countries and between the global south and global north. They could also build national and regional infrastructure for plant breeding programmes, and revise agricultural policies and subsidies to support more-diverse, sustainable and locally adapted food systems that are built on conventional and non-conventional crops.



Africa: sequence 100,000 species to safeguard biodiversity

But meaningful change — at scale — also requires scientists worldwide to re-evaluate their fundamental assumptions, and re-examine how they work with partners across cultures and power structures[26].

Take conservation. As noted by others, approaches to protecting ecosystems are often centred on the concept of some ideal 'wilderness' state. Yet nearly all landscapes have been actively managed and shaped by Indigenous people and local communities for millennia. Dominant narratives about 'pristine', people-free nature carry little, if any, regard for Indigenous and local ways of knowing, using and living in these landscapes[27].

Equity must be reflected in authorship lists, intellectual-property rights, patents and other forms of monetary and non-monetary sharing of benefits derived from research. But even the use of language in research, which reflects deeply entrenched biases and power imbalances, must be rethought[5].

Plant breeders and other stakeholders often point to the need for capacity building in low-income countries, for instance. The way the term is used overlooks the fact that people in these countries already have extensive knowledge and experience[4] — 'capacity' of their own, even if they lack infrastructure, equipment and some technical knowledge. Likewise, researchers continue to report the 'discovery' of a new species (including in

this journal), even when local communities have known about that species for hundreds or thousands of years.

Various initiatives are already helping to make biodiversity and agriculture research more accessible and inclusive. Since 2021, Africa's continental platform for open-access publishing has made studies conducted in Africa freely available, for the primary benefit of other researchers on the continent. Likewise, the African BioGenome Project aims to develop a major genomics resource in Africa to assist breeders and conservationists.

Around the world, botanical gardens, natural-history museums and private and public research and development organizations are trying to forge a better way forward.

At Kew, we are doing the same. My colleagues, together with around 210 collaborators from more than 40 countries, have compiled information about more than 7,000 plant species with documented uses as human food in an effort spanning more than 20 years[21,28]. We are digitizing our collections of some 7 million plant and 1.2 million fungal species, so that users anywhere in the world can have free access to high-resolution images and their associated data. And we are offering in-person courses in conservation science, seed-banking and biodiversity in more than 30 countries and territories worldwide, reaching many more people and regions online.

When I go back to Brazil, I always visit a local fruit and vegetable market. I spend hours chatting to the farmers, smelling and tasting the local produce and learning more about how they grow such an amazing diversity of edible plants. Their stories about what they are growing where — and most of all, their enthusiasm — give me hope that biodiversity is coming back to our lives, and will stay.

Nature **613**, 239-242 (2023)

doi: https://doi.org/10.1038/d41586-023-00021-4

References

1. Levis, C. *et al. Science* **355**, 925–931 (2017).

2. FAO. *Indigenous Peoples' Food Systems: Insights on Sustainability and Resilience from the Front Line of Climate Change* (FAO, Alliance of Bioversity International & CIAT, 2021).

3. Moran, K., King, S. R. & Carlson, T. J. *Annu. Rev. Anthropol.* **30**, 505–526 (2001).

4. Brondízio, E. S. *et al. Annu. Rev. Environ. Res.* **46**, 481–509 (2021).

5. Smith, L. T. *Decolonizing Methodologies: Research and Indigenous Peoples* (Bloomsbury, 2021).

6. Khoury, C. K. *et al. New Phytol*. **233**, 84–118 (2022).

7. Koch, A., Brierley, C., Maslin, M. M. & Lewis, S. L. *Quat. Sci. Rev.* **207**, 13–36 (2019).

8. McCann, J. C. *Maize and Grace: Africa's Encounter with a New World Crop, 1500–2000* (Harvard Univ. Press, 2005).

9. Das, S. & Lowe, M. *J. Nat. Sci. Collect.* **6**, 4–14 (2018).

10. Rausch, L. L. *et al. Conserv. Lett.* **12**, e12671 (2019).

11. Braga, A. R. C., de Rosso, V. V., Harayashiki, C. A. Y., Jimenez, P. C. & Castro, Í. B. *Nature Food* **1**, 312–314 (2020).

12. Pingali, P. L. *Proc. Natl Acad. Sci. USA* **109**, 12302–12308 (2012).

13. FAO. *The White/Wiphala Paper on Indigenous Peoples' Food Systems* (FAO, 2021).

14. Cabral, L., Pandey, P. & Xu, X. *Agric. Hum. Values* **39**, 249–267 (2022).

15. IPCC. *Climate Change and Land: An IPCC Special Report on Climate Change. Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems* (eds Shukla, P. R. *et al.*) (IPCC, 2019).

16. Díaz, S. *et al. Science* **366**, eaax3100 (2019).

17. Willett, W. *et al. Lancet* **393**, 447–492 (2019).

18. Pironon, S. *et al. Nature Clim. Change* **9**, 758–763 (2019).

19. Hamilton, A. C., Karamura, D. & Kakudidi, E. *Plant Divers*. **38**, 23–44 (2016).

20. Borrell, J. S. *et al. Ann. Bot.* **123**, 747–766 (2019).

21. Ulian, T. *et al. Plants People Planet* **2**, 421–445 (2020).

22. Gardner, E. M. *et al. Curr. Biol.* **32**, R511–R512 (2022).

23. Cantwell-Jones, A. *et al. Nature Plants* **8**, 225–232 (2022).

24. Vermeylen, S. *Local Environ.* **12**, 423–436 (2007).

25. Swiderska, K. *et al. Sustainability* **14**, 11311 (2022).

26. Ramírez-Castañeda, V. *et al. Proc. Natl Acad. Sci. USA* **119**, e2122667119 (2022).

27. Fletcher, M.-S., Hamilton, R., Dressler, W. & Palmer, L. *Proc. Natl Acad. Sci. USA* **118**, e2022218118 (2021).

28. Antonelli, A. *et al. State of the World's Plants and Fungi 2020* (Royal Botanic Gardens, Kew, 2020).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00021-4

| Section menu | Main menu |

- CORRESPONDENCE
- 10 January 2023

Use WHO antibody standards for emergency vaccine testing

- Valentina Bernasconi

The World Health Organization's International Standards for antibody assays enable researchers to compare results from different clinical trials of vaccines against a particular pathogen. Yet, despite the availability of such an antibody standard for SARS-CoV-2, none of the developers of the most advanced vaccine candidates used this standard when reporting early clinical data or when the vaccines were approved for emergency use.

Comparability of clinical-trial results improves the public-health value of vaccines. Regulators, clinical-trial sponsors, funders of vaccine development and ethics committees would all benefit from incorporating the WHO standards as a reporting requirement. Research papers should include them as a benchmark and report serology-assay data using the International System of Units for transparency.

Such standards are not always available at the start of a virus outbreak, as was the case for the Zika and 2014 Ebola outbreaks, so immune-response assays must be developed without such calibration. In that event, using standardized testing procedures and the same reagents can help.

Note also that the WHO issued a manual in April 2022 for preparing reference materials for use as secondary standards in antibody testing.

Nature **613**, 243 (2023)

doi: https://doi.org/10.1038/d41586-023-00007-2

Supported by 9 signatories (see Supplementary information for full list).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00007-2

Work

- **Sabbaticals are back: how to make the most of them** [11 Jan 2023]
  • Researchers reflect on past sabbatical breaks and give advice on taking one after the stresses of the pandemic.
- **The reproducibility issues that haunt health-care AI** [09 Jan 2023]
  • Health-care systems are rolling out artificial-intelligence tools for diagnosis and monitoring. But how reliable are the models?
- **Tracking ocean history through plankton ‘roadkill’** [09 Jan 2023]
  • Marine biogeochemist Clare Ostle tracks the stories that plankton tell from all corners of the globe.

- CAREER FEATURE
- 11 January 2023

Sabbaticals are back: how to make the most of them

Researchers reflect on past sabbatical breaks and give advice on taking one after the stresses of the pandemic.

- Amber Dance

Find a new job



Taking time out from academic responsibilities to try new experiences can boost creativity.Credit: Getty

In 2009, Brenda Wingfield completed the sequence of the pathogen *Fusarium circinatum* — the first fungal genome to be fully sequenced on the African continent — which causes local pine trees to wilt and ooze resin. But the fungal geneticist at the University of Pretoria in South Africa had a problem: she didn't know how to annotate and analyse a whole genome.

Wingfield found her solution half a world away in the United States, where *F. circinatum* is also a problem. She trained in these genomics techniques during a sabbatical at the Genome Center of the University of California, Davis.

"That was mind-blowing, learning genomics that I'd never done before," she recalls. Wingfield brought her new-found expertise back home to Pretoria, where she hosted a week-long "genome annotation jamboree".

The practice of academic sabbaticals began in 1880 at Harvard University in Cambridge, Massachusetts, and these professional breaks have become a perk of professorships around the world. Sabbaticals of up to one year are typically available to tenured faculty members every seven years, although policies vary by institution and nation. Some companies also offer sabbaticals or similar kinds of leave. Data on sabbatical rates are scarce, and not everyone is able to take advantage. Academics in some low- or middle-income nations, for instance, might face extra struggles with obtaining visas, crossing borders or the cost of living in nations with stronger currencies.

COVID-19 lockdowns stopped most international travel, and some scientists opted for at-home 'staybaticals', enjoying the rejuvenation and creativity of time away from teaching and administrative duties without the stress of travel. Now that borders are opening up again, conventional sabbaticals are once more an option.



This pandemic is not an extended sabbatical

“My sense is that we are almost back to ‘normal’ with regards to academic travel again,” says Wingfield. “I am aware of a number of people who are now on sabbatical or planning a sabbatical for 2023.”

Scientists use this time away from teaching and administrative responsibilities in various ways: to learn new skills or experimental systems; to write papers or books; to develop relationships and collaborations; or to take a career in a new direction. But those benefits come at some effort and cost — a successful sabbatical requires months or even years of planning to arrange a host, travel funds or a replacement salary, and to organize the logistics of uprooting a personal and professional life. Scientists might struggle to keep research going at their home institutions or to find the right schools for their children, for instance.

It can also be difficult for some faculty members to arrange a full year away. Rachel Spronken-Smith, a geographer at the University of Otago in Dunedin, New Zealand, notes that sabbaticals are encouraged at her institution, but researchers tend to take shorter trips. She is currently on an eight-month staybatical, in part because pandemic pressures at work left her too exhausted to plan a trip and too worried about getting ill while overseas. “Being on a staybatical has allowed me to refresh in every sense,” she says. She has moved her research forwards and updated her courses. She has also

taken short holidays to New Zealand destinations. Others might take a mini sabbatical lasting days or weeks (see 'Mini sabbaticals'). These options allow scientists to minimize financial, travel and logistical hurdles.

Mini sabbaticals

Even a short break yields rewards.

Not everyone has time to go away for six months or more. A mini sabbatical, lasting days or weeks, can give busy faculty members the chance to learn something new or extend their professional network. It can also be the right choice for junior researchers who haven't yet earned the right to a full sabbatical.

"One can come back from a week, two weeks or three weeks, perhaps with new skills, new ideas or new connections," says Michael Pillinger, a rheumatologist at New York University School of Medicine, who analysed mini sabbaticals offered by clinical institutions in a 2019 study.

Some scientists design their own trips. Others turn to official sabbatical programmes, such as that offered by the Northwestern-Argonne Institute of Science and Engineering in Evanston, Illinois. The programme allows faculty members at Northwestern University to spend at least three days a week at Argonne National Laboratory in nearby Lemont during one academic quarter.

Andreas Wächter, an industrial engineer at Northwestern University, took advantage of the programme in 2018. "You have the luxury of a huge chunk of time to concentrate on research," says Wächter. "You can really concentrate on establishing new relationships."

His new-found network included researchers at the Los Alamos National Laboratory in New Mexico, where Wächter went for a year-long sabbatical in 2019–20.

Even a short trip can provide benefits. Research shows that diverse experiences boost creativity — particularly when they involve other cultures[1,2]. And studies of sabbaticals find that researchers return home with

greater vitality, reduced stress levels and more career resources, especially if they travel abroad[3–5].

Home or away

Landing in a new place is part of the experience, says Wingfield. “You are removing yourself from your regular social circle,” she says. “That allows you to open yourself up to new ideas and to do things a bit differently.”

Wingfield credits her own career advancement and some of her most-cited papers to her five sabbaticals, taken between 1994 and 2017. Her work on *F. circinatum* and other fungal genomes[6], which was a complete change from her previous research, helped her to gain a research chair in fungal genomics.

For some scientists, sabbaticals are a matter of survival. Developmental biologist Yolanda Cruz retired in 2021 from Oberlin College in Ohio. Teaching there is a full-time activity, but faculty members are also expected to maintain a research programme. She had trained in mouse development, but felt she couldn’t compete with laboratories that had full-time research staff.

Cruz decided to focus on the less-crowded field of marsupial development. For her first sabbatical year, in 1992, she spent several months in Australia, learning about marsupial biology. She used the remainder in Oberlin, setting up her lab to study opossums. “That’s probably the best thing I ever did, career-wise,” says Cruz, because it opened up a new field of study in which she could be competitive.

Cruz arranged for a colleague at Oberlin to rent her house and car while she was gone. But the trip didn’t go perfectly. Cruz had bought an ancient Mitsubishi car for her six months in Australia, on the understanding that the dealer would buy it back when she left. But the dealer didn’t honour the arrangement. Cruz then had to drive around until she found a new buyer, who paid much less than she’d asked for.



Obstacle race: the barriers facing graduates who study abroad

Yet the experience didn't turn her off sabbaticals. She spent further stints in Australia and Europe, where she split her time between the labs of a former student and a past postdoctoral adviser, learning the latest molecular-biology techniques.

Going on sabbaticals has also been a crucial career move for Spronken-Smith. Because New Zealand is geographically isolated, she says, the university encourages faculty members to maintain overseas ties and collaborations. In a survey she published with her colleagues in 2015, she found that academic researchers in New Zealand considered sabbatical leave to be an important factor in career progression and promotions, second only to garnering grants[7].

Spronken-Smith has spent sabbaticals in Canada, the United Kingdom and the United States, but it hasn't always been possible to travel. Her children were young in 2000, when she was working at the University of Canterbury in Christchurch, New Zealand. The family had secured a coveted spot at the university's day-care centre that they didn't want to lose. So Spronken-Smith took a short trip to Europe for a conference and spent the rest of her sabbatical year at home.

To create some distance from her work in the geography department, she obtained space in the physics department. “The purpose was to find an office where nobody could find me,” she says. She enjoyed still having coffee with colleagues, while skipping departmental meetings.

The staybatical had a big influence on her career. As part of it, she earned a postgraduate diploma in university teaching. “That set me on a different career path” researching higher education, she says.

Break free

Heran Darwin, a microbiologist at New York University School of Medicine, needed a rejuvenating break when she took a sabbatical in 2015. Her lab building had been flooded by Hurricane Sandy in 2012, and she’d had to move her tuberculosis research lab four times before settling in a permanent space in 2018. It was stressful and disruptive.

She headed for the US West Coast, signing on as a visiting scientist at the University of California, Berkeley, where she learnt immunology and taught her hosts how to work with tuberculosis.



Many researchers report improved well-being and enhanced career focus after a sabbatical.Credit: Getty

Heading home, she felt a twinge of sadness, but was happy to return to her trainees and research with a new-found fluency in immunology along with new collaborators and projects. “I came back refreshed and motivated about science,” says Darwin.

Other researchers find that sabbaticals offer the perfect route to rev up their research after working in an administrative position. In 2012, materials scientist Gan Moog Chow stepped down from his last position in a nine-year stretch of administrative roles, ending his term as department head. “I felt I needed to take a sabbatical to renew and retool myself in research,” says Chow, who retired from the National University of Singapore at the end of 2022.

Chow had been working on chemical syntheses of magnetic thin films for data storage and nanoparticles for sensors. But he was ready for a change,

and wanted to tackle the structures of complex oxides, metal-containing compounds used for energy-efficient information storage.

To get up to speed, Chow spent five months in 2013 as a visiting scientist at Argonne National Laboratory in Lemont, Illinois, and another five in 2014 at Brookhaven National Laboratory in Upton, New York. He learnt about synchrotron X-ray techniques and scanning transmission electron microscopy, skills that he needed to delve into the oxide world. “Without the sabbatical, it would have been much harder for me to enter this highly competitive research field,” says Chow.

Hans Clevers also used a sabbatical year in 2015–16 to readjust to research after an administrative job. He’d just finished three years as president of the Royal Netherlands Academy of Arts and Sciences, advising the government, talking to reporters and travelling. But he was returning to his stem-cell research at the Hubrecht Institute in Utrecht, Netherlands.

Clevers designed a schedule of six consecutive sabbatical trips, spending six to eight weeks in each place. He visited Australia, Israel, New York, California, France and Hong Kong to give talks, speak to scientists and spread the word about the organoid technology developed in his lab. (Clevers is now head of pharmaceutical research and early development at the multinational drug firm Roche in Basel, Switzerland, which also grants sabbaticals to its employees.)

A couple of months was the perfect amount of time to get to know people and a place, Clevers says: “You don’t need a year.”

Plan ahead

Clevers estimates that he started planning his sabbatical about nine months in advance. Travelling professors usually have to apply for the leave well before they plan to depart, and many institutions require accomplishment reports to be submitted on return.

Most trips will require a host: a scientist in the new location who arranges for office or lab space and Internet and library access for the visiting

researcher. It isn't hard to find one if researchers plumb their professional networks, says Cruz.

Researchers must also find ways to finance their trip. Not all institutions cover salary for a full year. Wingfield uses grants to supplement funding for her sabbaticals, and notes that it can take two to three years to arrange the money (see 'Sabbatical funders'). Travellers might also have to pay for airfares and housing in a distant location while maintaining living quarters at home. Services such as SabbaticalHomes.com can help academics to find, rent or swap homes.

Sabbatical funders

Several organizations offer financial help for those seeking to visit other countries.

Scientists can cobble together funds for sabbaticals from a variety of sources, including monies that might be available at their home or host institutions. Be sure to ask around, and consider government and foundation grants, too. Some potential sources of money include:

• The Oppenheimer Memorial Trust funds sabbatical projects for academics at South African institutions.

• For non-Japanese researchers working in Japan, fellowships are available from the Japan Society for the Promotion of Science (Invitational Fellowships for Research in Japan) and the Matsumae International Foundation.

• The John Simon Guggenheim Memorial Foundation awards fellowships to citizens and permanent residents of the United States, Canada, Latin America and the Caribbean.

• The US Fulbright Visiting Scholar Program supports non-US scholars to conduct research at US institutions, and the Fulbright Scholar Program funds US academics to do research abroad.

• The Australian Academy of Science awards Selby Fellowships to visiting scientists for a public lecture tour.

With inflation currently high in many countries, exchange rates might also be a factor. Going to a nation that has a higher-value currency than at home can compound the financial challenge. While Wingfield was on her first sabbatical in the United States in 1994, the value of the South African rand dropped. She was travelling with her children, who wondered why she suddenly became more price-conscious when grocery shopping.

But Wingfield took the hit to her wallet in her stride. “It was good for my middle-class children to understand that they couldn’t always have those wonderful, coloured cereals,” she says.

For researchers with children, travelling abroad can also mean trying to sync school schedules in different countries. Wingfield’s son had to repeat a school year when he returned home after one sabbatical. But neither Wingfield nor her son felt that it was a great loss. He once told his mother with pride, “I’ve been to school on three continents.”

Scientists must also consider the students and postdocs in the labs at their home institutions. Many keep in contact through video calls. Wingfield often selects a senior graduate student or postdoc to keep an eye on the group. It also helps if students have more than one adviser to rely on.

Clevers checked in with his group in Utrecht between each of his mini sabbaticals, and says the team got on well without him, publishing several high-quality papers. “It was a bit of an eye-opener that I don’t need to be on top of everything,” he says.

Ultimately, a sabbatical can be whatever a researcher needs it to be. Cruz’s advice is to look at the breadth of possibilities that qualify as sabbatical projects, then ask, “What is the most amazing thing you can do?”

Nature **613**, 399-401 (2023)

doi: https://doi.org/10.1038/d41586-023-00022-3

References

1. Ritter, S. M. *et al. J. Exp. Soc. Psychol.* **48**, 961–964 (2012).

2. Maddux, W. W., Adam, H. & Galinsky, A. D. *Pers. Soc. Psychol. Bull.* **36**, 731–741 (2010).

3. Carraher, S. M., Crocitto, M. M. & Sullivan, S. *Car. Dev. Int.* **19**, 295–313 (2014).

4. Iravania, H. *Procedia Soc. Behav. Sci.* **15**, 3608–3615 (2011).

5. Davidson, O. B. *et al. J. Appl. Psychol.* **95**, 953–964 (2010).

6. van der Nest, M. A. *IMA Fungus* **5**, 473–486 (2014).

7. Smith, D., Spronken-Smith, R., Stringer, R. & Wilson, C. A. *Higher Ed. Res. Dev.* **35**, 589–603 (2015).

8. Pillinger, M. *et al. J. Clin. Transl. Sci.* **3**, 37–44 (2019).

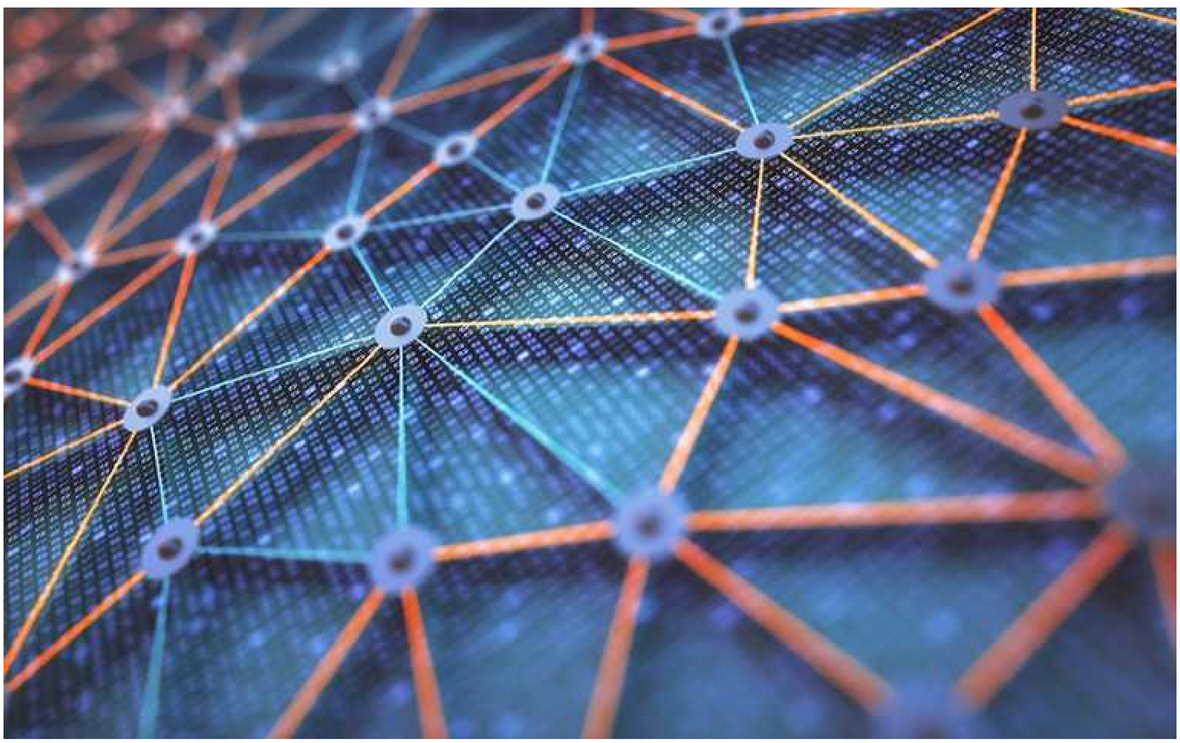
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00022-3

- TECHNOLOGY FEATURE
- 09 January 2023
- Correction 12 January 2023

The reproducibility issues that haunt health-care AI

Health-care systems are rolling out artificial-intelligence tools for diagnosis and monitoring. But how reliable are the models?

- Emily Sohn



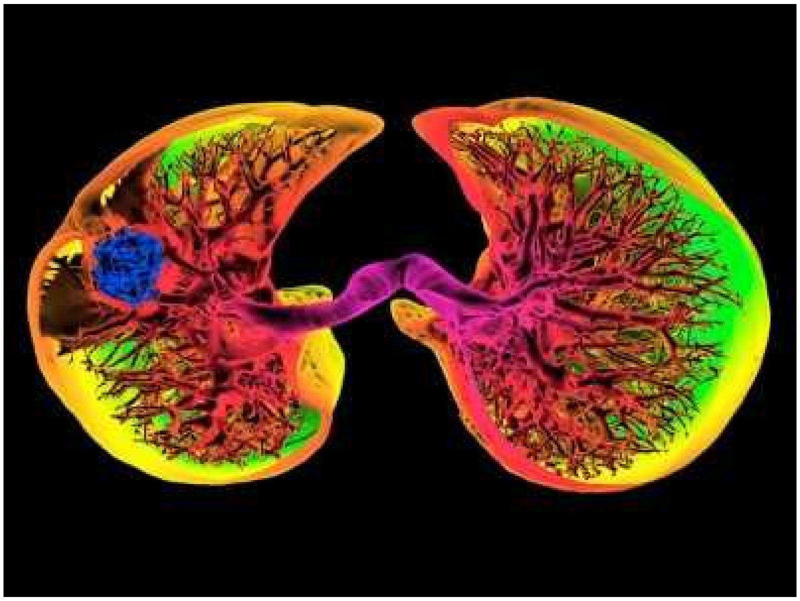
The use of artificial intelligence in medicine is growing rapidly.Credit: ktsimage/Getty

Each day, around 350 people in the United States die from lung cancer. Many of those deaths could be prevented by screening with low-dose computed tomography (CT) scans. But scanning millions of people would produce millions of images, and there aren't enough radiologists to do the work. Even if there were, specialists regularly disagree about whether images show cancer or not. The 2017 Kaggle Data Science Bowl set out to test whether machine-learning algorithms could fill the gap.

An online competition for automated lung cancer diagnosis, the Data Science Bowl provided chest CT scans from 1,397 patients to hundreds of teams, for the teams to develop and test their algorithms. At least five of the winning models demonstrated accuracy exceeding 90% at detecting lung nodules. But to be clinically useful, those algorithms would have to perform equally well on multiple data sets.

To test that, Kun-Hsing Yu, a data scientist at Harvard Medical School in Boston, Massachusetts, acquired the ten best-performing algorithms and challenged them on a subset of the data used in the original competition. On these data, the algorithms topped out at 60–70% accuracy, Yu says. In some cases, they were effectively coin tosses[1]. "Almost all of these award-winning models failed miserably," he says. "That was kind of surprising to us."

But maybe it shouldn't have been. The artificial-intelligence (AI) community faces a reproducibility crisis, says Sayash Kapoor, a PhD candidate in computer science at Princeton University in New Jersey. As part of his work on the limits of computational prediction, Kapoor discovered that reproducibility failures and pitfalls had been reported in 329 studies across 17 fields, including medicine. He and a colleague organized a one-day online workshop last July to discuss the subject, which attracted about 600 participants from 30 countries. The resulting videos have been viewed more than 5,000 times.



[Could machine learning fuel a reproducibility crisis in science?]

It's all part of a broader move towards increased reproducibility in health-care AI, including strategies such as greater algorithmic transparency and promoting checklists to avoid common errors.

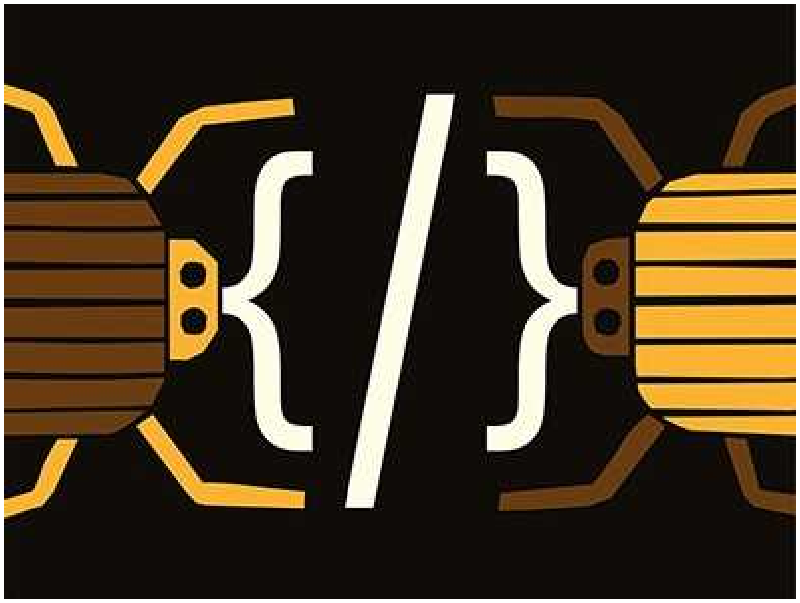
These improvements cannot come soon enough, says Casey Greene, a computational biologist at the University of Colorado School of Medicine in Aurora. "Given the exploding nature and how widely these things are being used," he says, "I think we need to get better more quickly than we are."

Big potential, high stakes

Algorithmic improvements, a surge in digital data and advances in computing power and performance have quickly boosted the potential of machine learning to accelerate diagnosis, guide treatment strategies, conduct pandemic surveillance and address other health topics, researchers say.

To be broadly applicable, an AI model needs to be reproducible, which means the code and data should be available and error-free, Kapoor says. But privacy issues, ethical concerns and regulatory hurdles have made reproducibility elusive in health-care AI, says Michael Roberts, who studies machine learning at the University of Cambridge, UK.

In a review[2] of 62 studies that used AI to diagnose COVID-19 from medical scans, Roberts and his colleagues found that none of the models was ready to be deployed clinically for use in diagnosing or predicting the prognosis of COVID-19, because of flaws such as biases in the data, methodology problems and reproducibility failures.



How to fix your scientific coding errors

Health-related machine-learning models perform particularly poorly on reproducibility measures relative to other machine-learning disciplines, researchers reported in a 2021 review[3] of more than 500 papers presented at machine-learning conferences between 2017 and 2019. Marzyeh Ghassemi, a computational-medicine researcher at the Massachusetts Institute of Technology (MIT) in Cambridge who led the review, found that a major issue is the relative scarcity of publicly available data sets in medicine. As a result, biases and inequities can become entrenched.

For example, if researchers train a drug-prescription model on data from physicians who prescribe medications more to one racial group than another, results could be skewed on the basis of what physicians do rather than what works, Greene says.

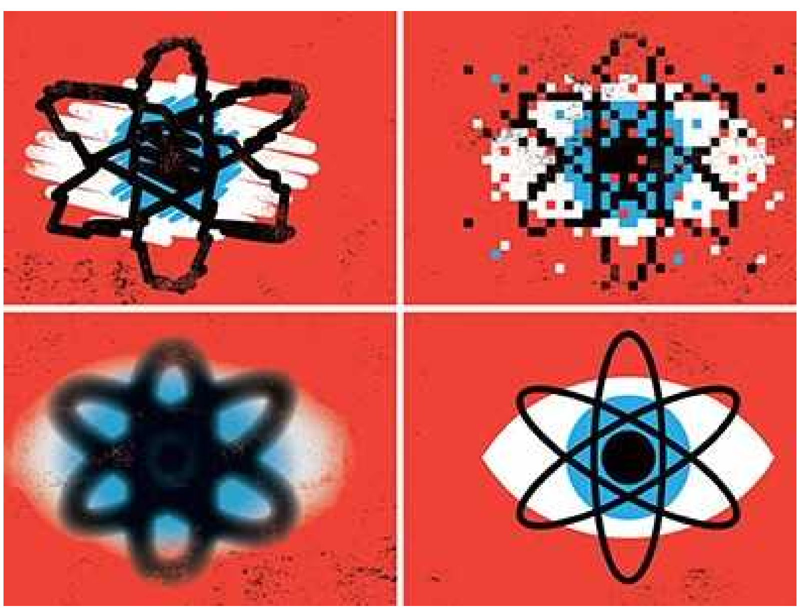
Another issue is data 'leakage': overlap between the data used to train a model and the data used to test it. These data sets should be completely

independent, Kapoor says. But medical databases can include entries for the same patient, duplications that scientists who use the data might not be aware of. The result could be an overly optimistic impression of performance, Kapoor says.

Septic shock

Despite these concerns, AI systems are already being used in the clinic. For instance, hundreds of US hospitals have implemented a model in their electronic health-record systems to flag early signs of sepsis, a systemic infection that accounts for more than 250,000 deaths in the United States each year. The tool, called the Epic Sepsis Model, was trained on 405,000 patient encounters at 3 health-care systems over a 3-year period, according to its creator Epic Systems, based in Verona, Wisconsin.

To evaluate it independently, researchers at the University of Michigan Medical School in Ann Arbor analysed 38,455 hospitalizations involving 27,697 people. The tool, they reported in 2021, produced a lot of false alarms, generating alerts on more than twice the number of people who actually had sepsis. And it failed to identify 67% of people who actually had sepsis[4]. (The company has since overhauled the models.)



Sharper signals: how machine learning is cleaning up microscopy images

Proprietary models make it hard to spot faulty algorithms, Greene says, and greater transparency could help to prevent them from becoming so widely deployed. “At the end of the day,” Greene says, “we have to ask, ‘Are we deploying a bunch of algorithms in practice that we can’t understand, for which we don’t know their biases, and that might create real harm for people?’”

Making models and data publicly available helps everyone, says Emma Lundberg, a bioengineer at Stanford University in California, who has applied machine learning to protein imaging. “Then someone could use it on their own data set and find, ‘Oh, it’s not working perfectly, so we’re going to tweak it’, and then that tweak is going to make it applicable elsewhere,” she says.

Positive moves

Scientists are increasingly moving in the right direction, Kapoor says, producing large data sets covering institutions, countries and populations, and that are open to all. Examples include the national biobanks of the United Kingdom and Japan, as well as the eICU Collaborative Research Database, which includes data associated with around 200,000 critical-care-unit admissions, made available by Amsterdam-based Philips Healthcare and the MIT Laboratory for Computational Physiology.

Ghassemi and her colleagues say that having even more options would add value. They have called for[3] the creation of standards for collecting data and reporting machine-learning studies, allowing participants to give consent to the use of their data, and adopting approaches that ensure rigorous and privacy-preserving analyses. For example, an effort called the Observational Medical Outcomes Partnership Common Data Model allows patient and treatment information to be collected in the same way across institutions. Something similar, the researchers wrote, could enhance machine-learning research in health care, too.

Eliminating data redundancy would also help, says Søren Brunak, a translational-disease systems biologist at the University of Copenhagen. In machine-learning studies that predict protein structures, he says, scientists

have had success in removing proteins from test sets that are too similar to proteins used in training sets. But in health-care studies, a database might include many similar individuals, which doesn't challenge the algorithm to develop insight beyond the most typical patients. "We need to work on the pedagogical side — what data are we actually showing to the algorithms — and be better at balancing that and making the data sets representative," Brunak says.



NatureTech

Widely used in health care, checklists provide a simple way to reduce technical issues and improve reproducibility, Kapoor suggests. In machine learning, checklists could help to ensure that researchers attend to the many small steps that need to be done correctly and in order, so that results are valid and reproducible, Kapoor says.

Multiple machine-learning checklists are already available, many spearheaded by the Equator Network, an international initiative to improve the reliability of health research. The TRIPOD checklist, for instance, includes 22 items to guide the reporting of studies of predictive health models. The Checklist for AI in Medical Imaging, or CLAIM, lists 42 items[5], including whether a study is retrospective or prospective, and how well the data match the intended use of the model.

In July 2022, Kapoor and colleagues published a list of 21 questions to help reduce data leakage. For example, if a model is being used to predict an outcome, the checklist advises researchers to confirm whether data in the training set pre-dates the test set, a sign that they are independent.

Although there is still much to do, growing dialogue around reproducibility in machine learning is encouraging and helps to counteract what has been a siloed state of research, researchers say. After the July online workshop, nearly 300 people joined a group on the online collaboration platform Slack to continue the discussion, Kapoor says. And at scientific conferences, reproducibility has become a frequent focus, Greene adds. “It used to be a small esoteric group of people who cared about reproducibility. Now it feels like people are asking questions, and conversations are moving forward. I would love for it to move forward faster, but at least it feels less like shouting into the void.”

Nature **613**, 402-403 (2023)

doi: https://doi.org/10.1038/d41586-023-00023-2

Updates & Corrections

- **Correction 12 January 2023**: An earlier version of this feature erroneously stated that Sayash Kapoor discovered reproducibility failures and pitfalls in 329 studies across 17 fields. In fact, those studies had themselves reported reproducibility failures and pitfalls.

References

1. Yu, K. H. *et al. J. Med. Internet Res.* **22**, e16709 (2020).
2. Roberts, M. *et al. Nature Mach. Intell.* **3**, 199–217 (2021).
3. McDermott, M. B. A. *et al. Sci. Transl. Med.* **13**, eabb1655 (2021).
4. Wong, A. *et al. JAMA Intern. Med.* **181**, 1065–1070 (2021).

5. Mongan, J., Moy, L. & Kahn, C. E. *Radiol. Artif. Intell.* **2**, e200029 (2020).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00023-2

- WHERE I WORK
- 09 January 2023

Tracking ocean history through plankton ‘roadkill’

Marine biogeochemist Clare Ostle tracks the stories that plankton tell from all corners of the globe.

- Kendall Powell



Clare Ostle is a marine biogeochemist at the Marine Biological Association in Plymouth, UK. Credit: Daniel Leal/AFP via Getty

In 1925, zoologist Sir Alister Hardy launched a survey of plankton while he was exploring Antarctica. Almost 100 years later, I'm continuing part of that same endeavour, called the Continuous Plankton Recorder (CPR) survey. Here, aboard our research vessel, the *Sepia*, I'm showing how the CPR works.

The device is a metre-long metal torpedo towed off a boat. Inside, two pieces of silk on a scroll sample the water, picking up plankton between them. We've got data and samples that date to 1931 from tows across the North Atlantic Ocean, which show what species of plankton are present, how they change across time and any new species that emerge.

Our taxonomists can identify 800 taxa of phytoplankton and zooplankton. They examine what I call plankton 'roadkill', after the organisms smash against the silk at 10 metres per second.



Collection: Fieldwork

We have routes that go from Canada to Japan, the United Kingdom to the Caribbean and even into the Arctic Ocean. With such a wide geographical and temporal span, we can look at large-scale spatial shifts in certain plankton species, such as those inhabiting cold-water zones contracting their ranges as warmer water moves polewards.

Container ships are ideal for consistently sampling from the same route every month. We can go on a limited budget, usually just biscuits and beer

for the crew. They enjoy knowing that ‘fishing for plankton’ is doing something interesting for science.

Because we store every sample, other researchers can find unexpected signals in the ocean later on. For example, a pioneering study from 2004 used CPR samples to show an increase in ocean microplastics. Another researcher tracked the bacteria that caused a 2018 cholera outbreak in British Columbia, Canada, with CPR samples. It’s exciting to be sitting on this database that someone with a different eye can pick up to tell untold stories.

Nature **613**, 406 (2023)

doi: https://doi.org/10.1038/d41586-023-00024-1

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-00024-1

| Section menu | Main menu |

Research

- ## An electric molecular motor [11 Jan 2023]
  • An electrically driven motor on the molecular scale based on [3]catenane is described, in which two cyclobis(paraquat-p-phenylene) rings operate by means of redox reactions, demonstrating highly unidirectional movement around a circular loop.
- ## Seasonal temperatures in West Antarctica during the Holocene [11 Jan 2023]
  • Analysis of a continuous record of water-isotope ratios from the West Antarctic Ice Sheet Divide ice core reveals a dominant role for annual maximum insolation in determining West Antarctic summer temperature during the Holocene.
- ## Periclase deforms more slowly than bridgmanite under mantle conditions [11 Jan 2023]
  • The authors model the creep of MgO periclase at lower mantle pressures and temperatures, finding that it deforms more slowly than bridgmanite at mantle strain rate.
- ## The molecular evolution of spermatogenesis across mammals [21 Dec 2022]
  • Evolutionary analyses of single-nucleus transcriptome data for testes from 11 species are reported, illuminating the molecular evolution of spermatogenesis and associated forces, and providing a resource for investigating the testis across mammals.
- ## Actin cytoskeleton and complex cell architecture in an Asgard archaeon [21 Dec 2022]
  • Culture and analysis of ‘Candidatus Lokiarchaeum ossiferum’—a member of the Asgard phylum—reveals an elaborate cell architecture with extensive membranous protrusions.
- ## Pathogen spillover driven by rapid changes in bat ecology [16 Nov 2022]
  • A study reveals how land-use change and climate interact to drive the spillover of a zoonotic virus, and identifies an ecological mechanism that prevents spillover.
- ## Integrated intracellular organization and its variations in human iPS cells [04 Jan 2023]
  • A dataset of 3D images from more than 200,000 human induced pluripotent stem cells is used to develop a framework to analyse cell shape and the location and organization of major intracellular structures.

- **A DNA methylation atlas of normal human cell types** [04 Jan 2023]
  • We describe a human DNA methylome atlas based on deep whole-genome bisulfite sequencing, allowing fragment-level analysis of cell-type-specific markers and providing an essential resource for studies of gene regulation and for deconvolution of cell mixtures and liquid biopsies.
- **Inheritance of paternal DNA damage by histone-mediated repair restriction** [21 Dec 2022]
  • In Caenorhabditis elegans, paternal exposure to ionizing radiation results in HIS-24 and HPL-1-dependent genome instability phenotypes, causing embryonic lethality in the offspring.
- **Structural basis of broad-spectrum β-lactam resistance in Staphylococcus aureus** [04 Jan 2023]
  • Cryo-electron microscopy structures of Staphylococcus aureus BlaR1 reveal dynamic signalling states regulating broad spectrum β-lactam antibiotic resistance through cleavage of the transcriptional repressor BlaI and induced expression of the β-lactamase blaZ and the β-lactam-resistant cell-wall transpeptidase mecA.

- Article
- Open Access
- Published: 11 January 2023

An electric molecular motor

- Long Zhang ,
- Yunyan Qiu,
- Wei-Guang Liu,
- Hongliang Chen,
- Dengke Shen,
- Bo Song,
- Kang Cai,
- Huang Wu,
- Yang Jiao,
- Yuanning Feng,
- James S. W. Seale,
- Cristian Pezzato,
- Jia Tian,
- Yu Tan,
- Xiao-Yang Chen,
- Qing-Hui Guo,
- Charlotte L. Stern,
- Douglas Philp,
- R. Dean Astumian ,
- William A. Goddard III &
- …
- J. Fraser Stoddart

Show authors

Nature **volume 613**, pages 280–286 (2023)

- 6197 Accesses

- 197 Altmetric
- Metrics details

Abstract

Macroscopic electric motors continue to have a large impact on almost every aspect of modern society. Consequently, the effort towards developing molecular motors[1,2,3] that can be driven by electricity could not be more timely. Here we describe an electric molecular motor based on a [3]catenane[4,5], in which two cyclobis(paraquat-*p*-phenylene)[6] ($CBPQT^{4+}$) rings are powered by electricity in solution to circumrotate unidirectionally around a 50-membered loop. The constitution of the loop ensures that both rings undergo highly (85%) unidirectional movement under the guidance of a flashing energy ratchet[7,8], whereas the interactions between the two rings give rise to a two-dimensional potential energy surface (PES) similar to that shown by F_OF_1 ATP synthase[9]. The unidirectionality is powered by an oscillating[10] voltage[11,12] or external modulation of the redox potential[13]. Initially, we focused our attention on the homologous [2]catenane, only to find that the kinetic asymmetry was insufficient to support unidirectional movement of the sole ring. Accordingly, we incorporated a second $CBPQT^{4+}$ ring to provide further symmetry breaking by interactions between the two mobile rings. This demonstration of electrically driven continual circumrotatory motion of two rings around a loop in a [3]catenane is free from the production of waste products and represents an important step towards surface-bound[14] electric molecular motors.

Main

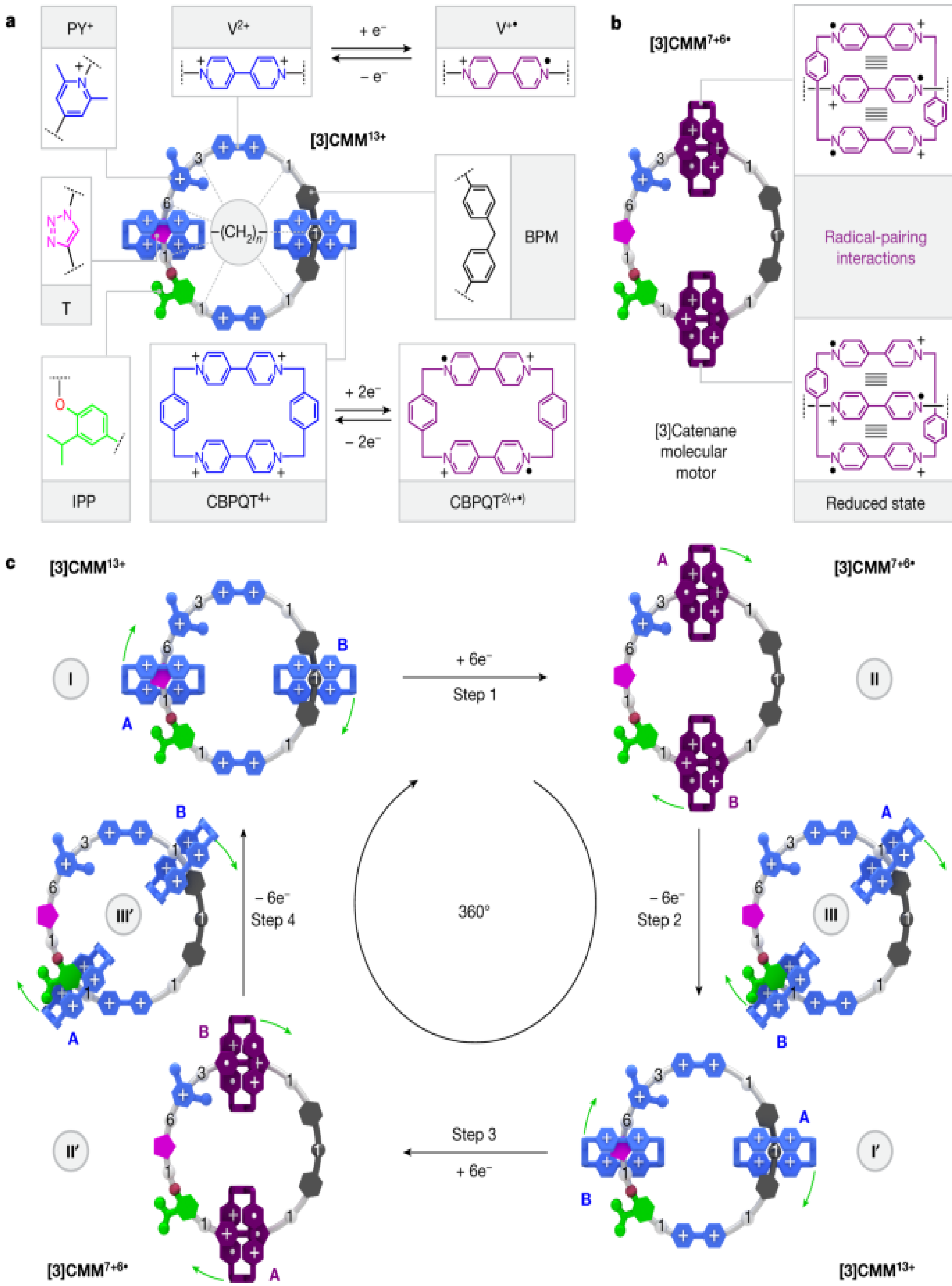
During the past 40 years, the design and synthesis[15,16] of artificial molecular machines have fostered the promise[17,18,19] of a technological revolution similar in magnitude to that arising from the development of macroscopic motors. To convert energy from an external source into unidirectional movement on a molecular scale[20,21], several artificial molecular machines, including[3,22,23] rotaxane-based linear motors[5,8,24,25,26,27] and catenane-based

rotary motors[4,5,28,29], have been designed, synthesized and shown to work in the presence of light[4,24,30,31,32] and chemical fuels[5,25,29,33,34]. Although it has been demonstrated[35,36,37] that single-molecule motors can be powered using tunnelling currents under ultrahigh vacuum on surfaces, examples of electrically powered catenane rotary motors operating in solution are not known to the best of our knowledge. Here we report the design, synthesis and operation of a redox-driven rotary motor based on a [3]catenane in which two rings can be powered by electricity to rotate unidirectionally around a loop.

The [3]catenane molecular motor **[3]CMM** comprises (Fig. 1a) two $CBPQT^{4+}$ (ref. [6]) rings encircling a 50-membered loop. A bis(4-methylenephenyl)methane (BPM) unit separates two viologen (V^{2+}) units, which are preordained, on reduction to their $V^{+\bullet}$ reduced radical cationic states, to serve as recognition sites[38] for reduced $CBPQT^{2(+\bullet)}$ rings. The remainder of the loop is composed of a chain containing 11 methylene groups and one oxygen atom, intercepted along its length by an isopropylphenylene (IPP) steric barrier, a triazole (T) ring, which is generated during the final ring closure of the loop to give the [3]catenane, and a 2,6-dimethylpyridinium (PY^{+}) Coulombic barrier. As a consequence of the redox properties of the $V^{2+/+\bullet}$ units in the loop and the two $CBPQT^{4+/2(+\bullet)}$ rings, the oxidized state **$[3]CMM^{13+}$** can be converted (Fig. 1b) into a reduced state **$[3]CMM^{7+6\bullet}$**, accompanied by movement of the two $CBPQT^{2(+\bullet)}$ rings, so as to encircle the $V^{+\bullet}$ units in the loop by virtue of radical-pairing interactions[39]. This design received its inspiration from previous investigations[12,25,40,41] on the redox-driven rotaxane-based molecular pumps, in which a pumping cassette[26,42,43], comprising PY^{+}–V^{2+}–IPP, serves as an active one-way gate to transport a $CBPQT^{4+}$ ring from the bulk solution onto a collecting chain after every redox cycle. In comparison with these linear rotaxane-based pumps, catenane-based motors are expected to undergo continuous unidirectional circumrotation of one ring about the other, as long as there is a power source. The obvious design in which the two ends of a rotaxane-based pump are linked to form a [2]catenane gives rise to a constitution in which there is no probe of unidirectionality. Consequently, we developed the loop and the [3]catenane

to investigate the interactions between and the unidirectionality of the two mobile rings.

Fig. 1: Design and working mechanism of the [3]catenane molecular motor [3]CMM.



a, Graphical representations with key structural fragments for the oxidized state of the [3]catenane molecular motor **[3]CMM^{13+}**. The

cyclobis(paraquat-*p*-phenylene) rings, the bisradical dicationic states of cyclobis(paraquat-*p*-phenylene), the viologens, the radical cationic states of the viologens, the bis(4-methylenephenyl)methane, the isopropylphenylene, the triazole and the 2,6-dimethypyridinium units are labelled as $CBPQT^{4+}$, $CBPQT^{2(+\bullet)}$, V^{2+}, $V^{+\bullet}$, BPM, IPP, T and PY^{+}, respectively. **b**, Graphical representations for the reduced state of the [3]catenane molecular motor **[3]CMM$^{7+6\bullet}$** with key superstructural formulas showing the radical-pairing interactions between the $CBPQT^{2(+\bullet)}$ rings and the $V^{+\bullet}$ units. Positive charges are balanced by PF_6^{-} counterions, which are omitted for the sake of clarity. **c**, The redox operation of **[3]CMM$^{13+/7+6\bullet}$** demonstrating the unidirectional rotary motion of the two $CBPQT^{4+}$/$CBPQT^{2+\bullet}$ rings. In state I, $[CBPQT\text{-}A]^{4+}$ and $[CBPQT\text{-}B]^{4+}$ are positioned around the T and BPM units, respectively. Reduction of the V^{2+} units and the $CBPQT^{4+}$ rings by the injection (step 1) of six electrons in total triggers both rings to undergo a clockwise rotation, leading to the formation (state II) of the reduced state **[3]CMM$^{7+6\bullet}$**. Subsequent oxidation by the removal (step 2) of six electrons restores the Coulombic repulsion between the two rings and the loop, obliging $[CBPQT\text{-}B]^{4+}$ to thread over (state III) the steric barrier (IPP) under thermal activation and eventually encircle T, whereas $[CBPQT\text{-}A]^{4+}$ finds itself threaded around BPM, thus completing a 180° positional exchange between the two rings shown in state I′. A second redox cycle (steps 3 and 4) resets the system back to state I after the accomplishment of another 180° positional exchange between the two rings.

Initially, we studied (Supplementary Scheme 5) the homologous [2]catenane **[2]C^{9+}**, in which only one $CBPQT^{4+}$ ring encircles the BPM unit on the loop. No directional motion was observed (Supplementary Fig. 35) in the [2]catenane. Instead, the single $CBPQT^{4+/2(+\bullet)}$ ring switches between one of the $V^{2+/+\bullet}$ units and the BPM unit during a redox cycle. To gain further insight into the switching behaviour of the [2]catenane, quantum mechanical calculations were undertaken (Supplementary Information Section 6.1). Several features were shown (Extended Data Figs. 1 and 2) in the calculated PESs of the oxidized and reduced forms of the [2]catenane. (1) The energy of the electrostatic barriers arising from the PY^{+} and $V^{2+/+\bullet}$ units, as well as the interaction energies between the $CBPQT^{4+/2(+\bullet)}$ ring and the $V^{2+/+\bullet}$ units, can be modulated simultaneously by switching between the oxidized and

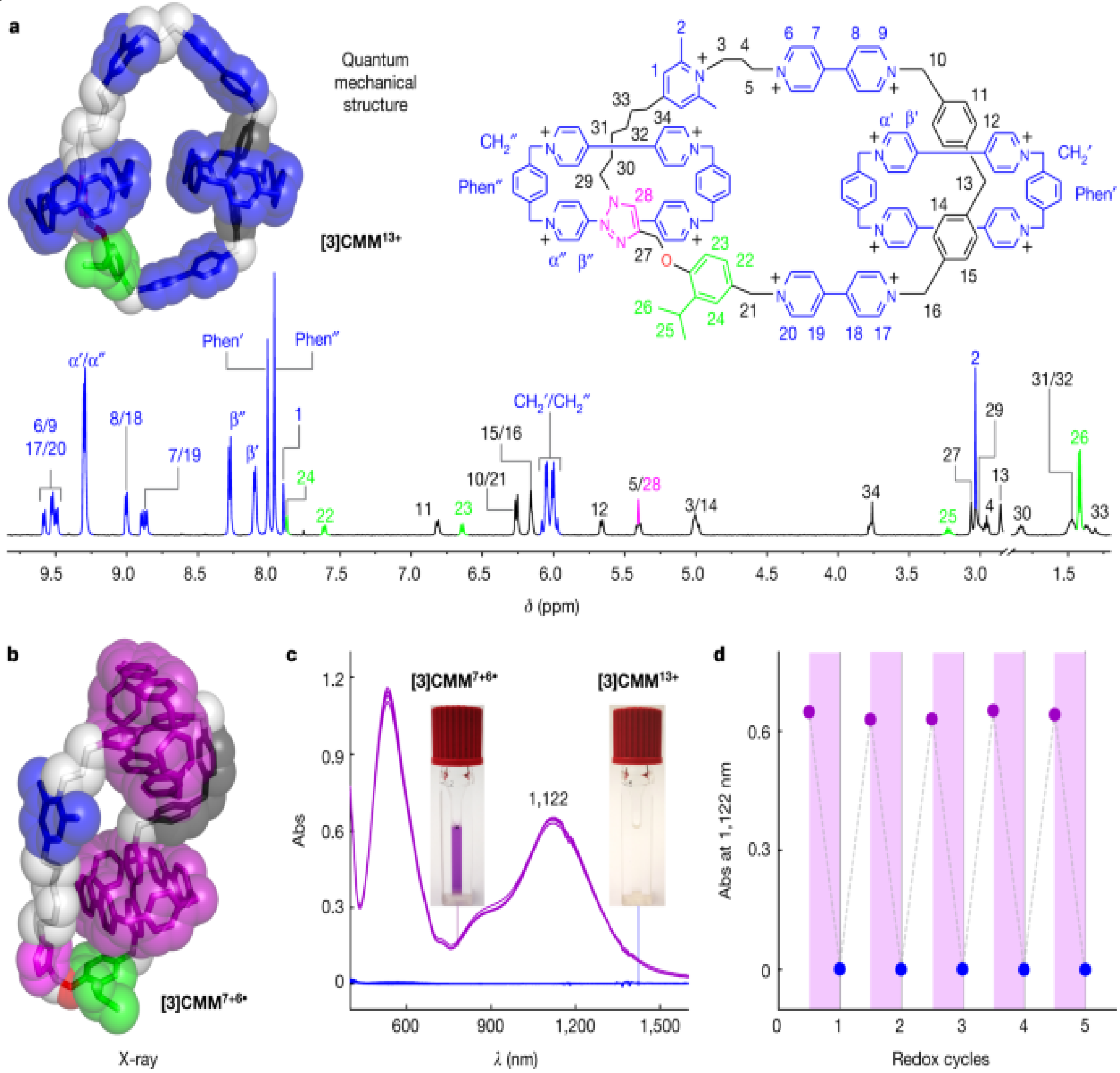
reduced states of the [2]catenane. (2) The height of the steric barrier imposed by the IPP unit is independent of the redox state and is large enough ($\Delta E > 25$ kcal mol^{-1}) to preclude the passage of the $CBPQT^{4+/2(+\bullet)}$ ring over the IPP unit. (3) An energy well, close to the centre of the BPM unit, traps the $CBPQT^{4+}$ ring under oxidizing conditions, thereby preventing directional movement on relevant timescales.

These results motivated our investigation of the [3]catenane. When two $CBPQT^{4+}$ rings encircle the loop under oxidizing conditions, one ring occupies the well provided by the BPM unit, thus requiring the passage of the other ring over the IPP unit. This combination of two mechanically interlocked[44] $CBPQT^{4+/2(+\bullet)}$ rings, a switchable barrier, associated with the PY^{+} unit, two switchable viologen ($V^{2+/+\bullet}$) recognition sites and a steric barrier (IPP) is key to inducing unidirectional motion (Fig. 1c) of the two rings with respect to the loop. Let us consider the mechanism of operation (Fig. 1c). To aid the discussion, the $CBPQT^{4+}$ rings are given the descriptors $[CBPQT\text{-}A]^{4+}$ and $[CBPQT\text{-}B]^{4+}$ (Fig. 1c, I). At the outset, on account of repulsive Coulombic forces, $[CBPQT\text{-}A]^{4+}$ resides around the T unit, whereas $[CBPQT\text{-}B]^{4+}$ encircles the BPM unit. On reduction, six electrons are injected (Fig. 1c, step 1) into the [3]catenane, resulting in both $CBPQT^{4+}$ rings being reduced to $CBPQT^{2(+\bullet)}$ and the V^{2+} units to $V^{+\bullet}$. The Coulombic interactions are decreased,allowing the attractive radical–radical interactions to dominate (Fig. 1c, II). As indicated (Extended Data Fig. 3) by the PESs of the reduced [3]catenane, the relative energy barrier for a $CBPQT^{2(+\bullet)}$ ring to traverse the PY^{+} unit is much lower than the barrier for the same ring to traverse the IPP unit. It follows that $[CBPQT\text{-}A]^{2(+\bullet)}$ passes over the PY^{+} unit to thread on to $V^{+\bullet}$ in a clockwise direction and $[CBPQT\text{-}B]^{2(+\bullet)}$ moves concomitantly to occupy the $V^{+\bullet}$ unit that neighbours the IPP unit. In this overall reduced state, the stable **[3]CMM**$^{\mathbf{7+6\bullet}}$ form is the one in which both $CBPQT^{2(+\bullet)}$ rings encircle[38] the $V^{+\bullet}$ binding sites. Subsequent oxidation (Fig. 1c, step 2) leads (Extended Data Fig. 4) to a biased Brownian motion[7,45] (Fig. 1c, III) to install the $[CBPQT\text{-}A]^{4+}$ ring on the BPM unit and the $[CBPQT\text{-}B]^{4+}$ ring on the T unit. Overall, one redox cycle triggers the positional exchange (Fig. 1c, I′) between the two $CBPQT^{4+}$ rings and completes a 180° unidirectional rotation. A subsequent redox cycle (Fig. 1c, steps 3 and 4) brings the rings back to their initial starting positions and

finishes a full 360° clockwise circumrotation around the loop. Steps 3 and 4 are identical to steps 1 and 2, respectively, but with the roles of the [CBPQT-A]$^{4+/2(+•)}$ and [CBPQT-B]$^{4+/2(+•)}$ rings reversed. The use of the term 'clockwise' means that each CBPQT^{4+} ring visits (Extended Data Fig. 5) the substituents on the loop in the order T/PY^{+}/IPP, rather than the other way around.

The [3]catenane was synthesized (Supplementary Scheme 6) by using radical templation[41] leading, in the first instance, to the formation of an intermediate pseudo[3]rotaxane before the final loop closure. The oxidized state **[3]CMM^{13+}**, which was isolated as its PF_6^- salt, was fully characterized (Fig. 2a) by ^{1}H nuclear magnetic resonance (NMR) spectroscopy. The positions of each CBPQT^{4+} ring on the loop were identified by changes in the chemical shifts of proton resonances on the BPM unit (H-13 and H-14) and on the triazole ring (H-28) as a result of the strong shielding by the encircling CBPQT^{4+} rings. The reduced state **[3]CMM$^{7+6•}$** of the [3]catenane, which was produced on the addition of 6.0 molar equivalents of cobaltocene[26] (Cp_2Co), or an excess of Zn dust[25], into an MeCN solution of **[3]CMM**•13PF_6 was accompanied by an instantaneous change from colourless to dark purple. The formation of **[3]CMM$^{7+6•}$** can also be followed (Supplementary Fig. 48) by visible/near infrared (Vis/NIR) spectroscopy with a broad absorption band centred on 1,122 nm, which is characteristic[38] of trisradical interactions. The solid-state structure of **[3]CMM$^{7+6•}$** was characterized by X-ray crystallography (Fig. 2b and Supplementary Information Section 7), which confirms the presence of the radical-pairing interactions (Supplementary Fig. 44) between the reduced CBPQT$^{2(+•)}$ rings and the V$^{+•}$ units. On addition of $NOPF_6$, reoxidation of **[3]CMM$^{7+6•}$**, which occurs within seconds, is accompanied by the MeCN solution reverting back to being colourless. The reversible switching between the oxidized and reduced states of the [3]catenane was also investigated (Extended Data Fig. 6 and Supplementary Information Section 8) by cyclic voltammetry, which was carried out on **[3]CMM**•13PF_6 in an MeCN solution. These investigations demonstrated that these two stable states of the [3]catenane can be interconverted using either chemical or electrochemical stimuli.

Fig. 2: Characterization of the redox state and electrically driven operation of the [3]catenane molecular motor [3]CMM.



a, Structural formula for the oxidized state **[3]CMM^{13+}** with an optimized quantum mechanical model structure (M06-2X/6-31G* basis set, in tubular with superimposed space-filling representation) and the ^{1}H NMR spectrum (500 MHz, CD_3COCD_3, 298 K), in which all the proton assignments are labelled. **b**, X-ray single-crystal structure of the reduced state **[3]CMM$^{7+6\bullet}$** depicted by tubular with superimposed space-filling representations. Solvent molecules, counterions and hydrogen atoms are omitted for the sake of clarity. **c**, Vis/NIR spectra of the reduced state **[3]CMM$^{7+6\bullet}$** (purple) and the oxidized state **[3]CMM^{13+}** (blue) during the electrically driven operation of

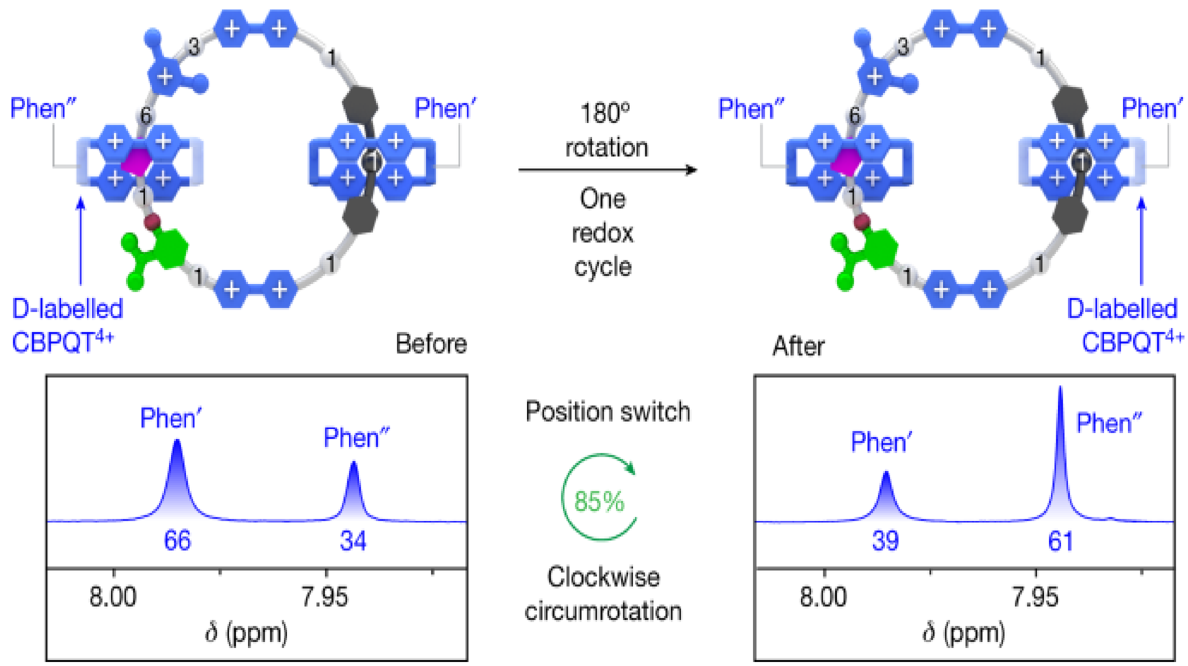
the molecular motor. Conditions: **[3]CMM** (30 μM), MeCN solution with $TBAPF_6$ (0.1 M) as the supporting electrolyte, reduction potential −0.5 V (versus Ag/AgCl) for 10 min, oxidation potential +0.7 V (versus Ag/AgCl) for 15 min. Insets are photographs of the two solutions in the oxidized (colourless) and reduced (purple) states. **d**, Absorption intensities of **[3]CMM$^{7+6\bullet}$** (purple) and **[3]CMM^{13+}** (blue) at 1,122 nm, showing the reversible switching between the two redox states during each cycle. Abs, absorption; ppm, parts per million.

With the reversible redox switching of the [3]catenane demonstrated, we turned our attention to using electricity to power the molecular motor by controlled potential electrolysis[11,26,41] (CPE). The electrically driven operation (Supplementary Information Section 10) of the molecular motor was accomplished in an electrochemical cell (Extended Data Fig. 7 and Supplementary Video 1) by alternating between two constant potentials (−0.5 V for reduction and +0.7 V for oxidation). The Vis/NIR spectrum of the reduced [3]catenane (30 μM) in MeCN was found to be in excellent agreement (Supplementary Fig. 49) with that obtained from the chemically driven operation with the same concentration (30 μM), confirming the validity of the CPE protocol. We repeated this protocol on the [3]catenane. The whole process was monitored (Fig. 2c) by Vis/NIR spectroscopy. The change in absorbance at 1,122 nm demonstrated (Fig. 2d) that the continuous operation of the motor can be repeated at least five times without any substantial loss in reversibility. The ^{1}H NMR spectrum (Supplementary Fig. 51) of the final oxidized motor molecule demonstrated recovery of more than 95% of the [3]catenane with negligible degradation, an observation that is commensurate with robust and reliable operation.

By simply accepting the present design of the [3]catenane, it is not possible to investigate the unidirectionality of the motor by determining the positions of the components after one 180° circumrotation on account of the fact that the two $CBPQT^{4+}$ rings are constitutionally identical. To address this conundrum, [D_{16}]-**$CBPQT^{4+}$** (Supplementary Scheme 7), in which the *p*-xylylene units are deuterium-labelled, was introduced (Extended Data Fig. 8 and Supplementary Scheme 9) during the synthesis of a deuterium-labelled [3]catenane [D_n]-**[3]CMM^{13+}**. As the $CBPQT^{2(+\bullet)}$ ring does not pass over the IPP unit under reducing conditions, there is only one gate (PY^+ unit) for

the CBPQT$^{2(+•)}$ rings threading to form the intermediate pseudo[3]rotaxane. Therefore, the distribution of undeuterated CBPQT^{4+} rings on the T and BPM units can be influenced by sequentially adding [D$_{16}$]-**CBPQT$^{2(+•)}$** and **CBPQT$^{2(+•)}$**. The [3]catenane [D$_n$]-**[3]CMM^{13+}** was obtained as a mixture of four isotopologues, including [D$_0$]-**[3]CMM^{13+}** and [D$_{32}$]-**[3]CMM^{13+}**, in addition to two co-constitutional isomers of [D$_{16}$]-**[3]CMM^{13+}**. Their presence was confirmed by high-resolution electrospray ionization mass spectrometry (Supplementary Fig. 55) and ^{1}H NMR spectroscopy (Supplementary Fig. 56). According to the integration of probe resonances for the H-phen′ and H-phen″ in the ^{1}H NMR spectra, the ratio of the undeuterated CBPQT^{4+} rings on the T and BPM units was found to be biased (Supplementary Figs. 56–59), allowing the system as a whole to be monitored. The deuterated motor [D$_n$]-**[3]CMM^{13+}** was subjected to CPE operation to test for the unidirectionality of circumrotation driven by electricity. The unidirectional circumrotation of the rings causes (Fig. 3) co-constitutional exchanges after one redox cycle to be quantified (Extended Data Fig. 9) on the basis of different ratios of integrated values of proton resonances (H-phen′ and H-phen″) in the ^{1}H NMR spectrum. From the ^{1}H NMR integrations (Fig. 3), the directionality of the molecular motor was calculated (Extended Data Fig. 10 and Supplementary Information Section 12) to be 85%, that is, 85% of the molecular motors complete a 180° unidirectional (clockwise) rotation in one redox cycle. Several chemically driven operations of the molecular motor were also performed (Supplementary Figs. 58–61), each providing comparable unidirectionalities of about 85% after one redox cycle.

Fig. 3: Measurement of the unidirectionality.

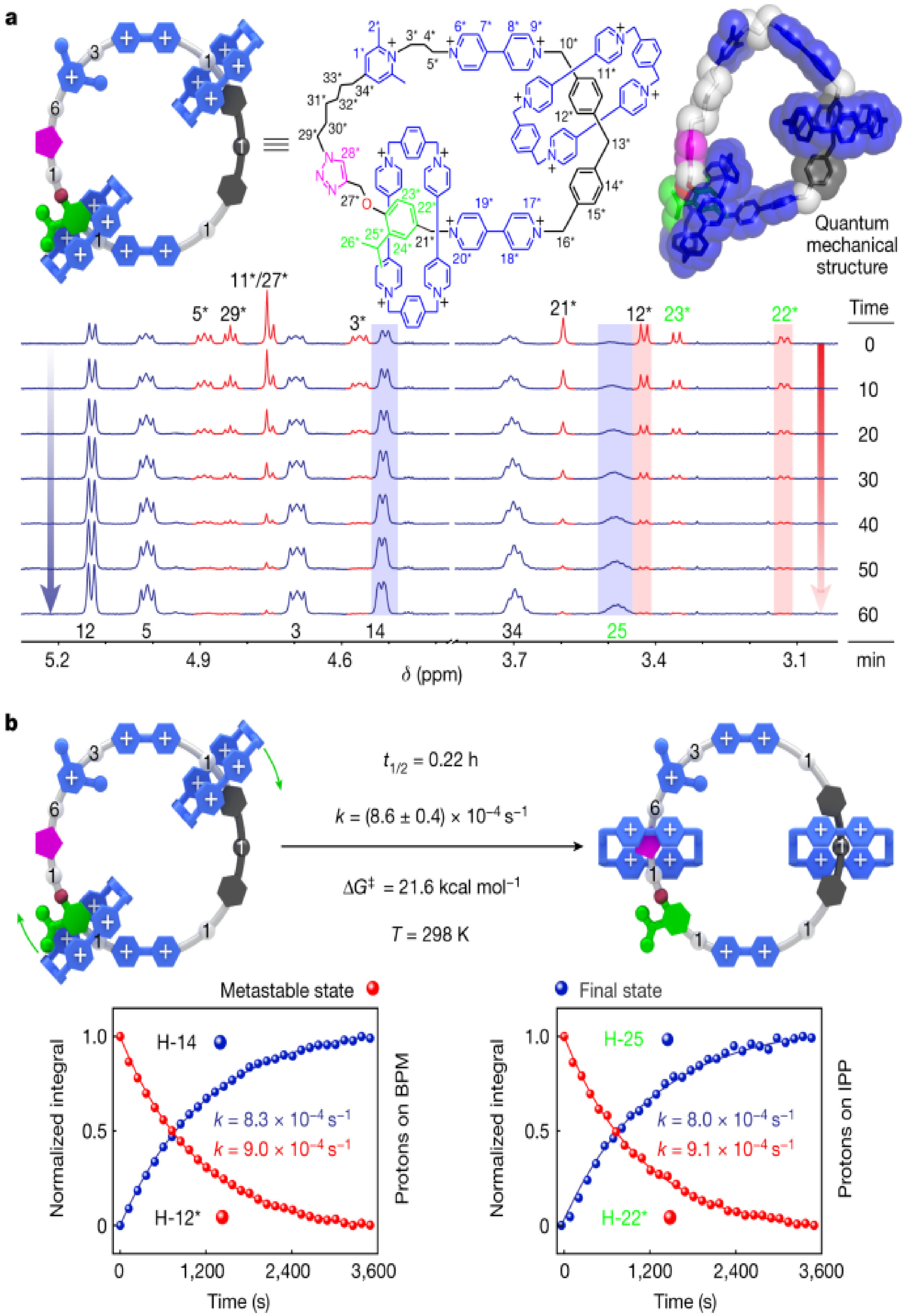


Top: graphical representation of the positional exchange of the deuterium-labelled $CBPQT^{4+}$ and $CBPQT^{4+}$ rings on the loop after one redox cycle. Bottom: partial ^{1}H NMR (600 MHz, CD_3COCD_3, 298 K) spectra of [D_n]-**[3]CMM**•$13PF_6$ with proton assignments before (left) and after (right) one electrically driven redox cycle. Numbers under peaks indicate relative integrals. ppm, parts per million.

A metastable state is observed (Fig. 4a, Supplementary Section 13 and Supplementary Fig. 66) during the oxidation process, demonstrating the unidirectional movement of both rings around the loop in the [3]catenane. Comparison (Supplementary Fig. 65) between the ^{1}H NMR spectrum of **[3]CMM**•$13PF_6$ and that of the metastable state shows the locations on the loop of two rings. Large upfield shifts are observed (Supplementary Fig. 65) for the resonances associated with the protons belonging to the BPM (H-12*) and the IPP (H-22*) units, whereas resonances for protons H-14* and H-15* are shifted downfield, indicating (Fig. 4a) that one of the two $CBPQT^{4+}$ rings moves away from the V^{2+} unit and is located asymmetrically on the BPM unit on account of net Coulombic repulsion, whereas the other ring is poised to mount the IPP steric barrier. The positions of the two rings on the loop indicate that the circumrotation towards the metastable state is in the clockwise direction. The subsequent relaxation by clockwise rotation (Extended Data Fig. 10) to the stable

oxidized state is favoured. By contrast, the counterclockwise motion going from the metastable state III to the stable oxidized state I would require both of the $CBPQT^{4+}$ rings to pass over the V^{2+} units, which is kinetically much less likely. The appearance of the metastable state and subsequent thermally activated relaxation to the stable oxidized [3]catenane was monitored (Fig. 4a and Supplementary Fig. 76) by ^{1}H NMR spectroscopy at 298 K in CD_3CN. Kinetic analysis (Fig. 4b and Supplementary Figs. 77 and 78) shows that this co-conformational rearrangement follows first-order kinetics at 298 K with an average rate constant k of $(8.6 \pm 0.4) \times 10^{-4}\ s^{-1}$, corresponding to an energy of activation ($\Delta G^{\ddagger}$) of 21.6 kcal mol^{-1}. These values are in good agreement with the calculated (Extended Data Fig. 4) value ($\Delta E^{\ddagger} = 20.1$ kcal mol^{-1}). The nearly identical rate constants (Supplementary Figs. 77 and 78) associated with relaxation of the metastable state of each ring suggests that, on oxidation, the two rings undergo unidirectional circumrotation (Supplementary Video 2) on account of the electrostatic interactions arising from the nanoconfinement[46] present in the mechanically interlocked [3]catenane. This mechanism is different from previously reported[4,5] [3]catenane-based molecular motors in that our electric molecular motor is driven by a single stimulus, namely the continuous oscillation of redox potential, that is, an applied direct current voltage.

Fig. 4: Metastable state in the redox cycle.



a
Quantum mechanical structure
11*/27*
5*
29*
3*
21*
12*
23*
22*
Time
0
10
20
30
40
50
60
12
5
3
14
34
25
5.2
4.9
4.6
3.7
3.4
3.1
δ (ppm)
min
b
$t_{1/2}$ = 0.22 h
k = (8.6 ± 0.4) × 10^{-4} s^{-1}
$\Delta G^{\ddagger}$ = 21.6 kcal mol^{-1}
T = 298 K
Metastable state
Final state
H-14
H-12*
k = 8.3 × 10^{-4} s^{-1}
k = 9.0 × 10^{-4} s^{-1}
Protons on BPM
H-25
H-22*
k = 8.0 × 10^{-4} s^{-1}
k = 9.1 × 10^{-4} s^{-1}
Protons on IPP
Normalized integral
Time (s)
0
0.5
1.0
1,200
2,400
3,600

a, Top, graphical representation and structural formula of the metastable state with an optimized quantum mechanical model structure (M06-2X/6-31G* basis set, in tubular with superimposed space-filling representation). Bottom, partial ^{1}H NMR (600 MHz, CD_3CN, 298 K) spectra of **[3]CMM**•$13PF_6$ measured over time (0–60 min) immediately after a cycle of reduction (Cp_2Co) and reoxidation ($NOPF_6$), with proton assignments labelled at the top and bottom of the spectra. The proton resonances attributable to the metastable state are labelled with an asterisk. **b**, Top, thermal relaxation associated with the co-conformational rearrangement from the metastable state to the reoxidized state. The activation energy barrier $\Delta G^{\ddagger}$ of 21.6 kcal mol^{-1} was determined using the Eyring equation ($k=\frac{{k}_{{\rm{B}}}T}{h}{{\rm{e}}}^{\frac{{-\Delta G}^{\ddagger}}{RT}}$), in which k is the reaction rate constant, T is the absolute temperature, R is the gas constant, $k_{\rm B}$ is the Boltzmann constant and h is the Planck constant. Bottom, plot of the changes in the normalized integral of protons on the BPM (H-12* and H-14) and IPP (H-22* and H-25) units with time at 298 K during the transformation from the metastable to the reoxidized state, as well as the fitting curves of these data according to the first-order kinetic model. ppm, parts per million.

We often think hierarchically about components of macroscopic machines, in which one component serves to drive another. This kind of hierarchy is impossible in molecular machines, in which movements of components obey the principle of microscopic reversibility. By contrast, the emergent interactions between the two rings in the electric molecular motor that allow for directional motion are reciprocal, but in conjunction with the kinetic asymmetry introduced into the structure of the molecule, they provide a mechanism by which time-symmetric switching between reducing and oxidizing conditions can drive unidirectional rotation. It may well turn out that these noncovalent interactions under nanoconfinement play the role of what is described as ‘gearing’ in macroscopic machines.

This electrically driven molecular motor exhibits several attractive features that are expected to play an important role in the subsequent development of artificial molecular machines. The [3]catenane can be synthesized in only four steps based on the previously described protocols[40,41]. Moreover, no covalent bonds need to be broken during the unidirectional circumrotation of

the two rings. The timescale for the completion of two redox cycles involving the unidirectional circumrotation of the two small rings through 360° with respect to the loop is only a few minutes. Furthermore, the co-constitution of the [3]catenane is compatible with attachment to the surface of an electrode by chemical modification of one of the two small rings, allowing spatially directed rotation with respect to a fixed frame of reference, thereby transducing electrical into mechanical energy at a surface.

The energy ratchet mechanism, on which this electric molecular motor is based, is powered by redox control of the two-dimensional potential energy landscape of the [3]catenane. The energy landscape is sculpted by the design of the constitution of the loop such that externally applied oscillation of the redox potential provides the energy required to harness Brownian motion of the molecular system to achieve unidirectionality. Our design represents a chemical rather than mechanical engineering approach to the development of electric molecular motors[47].

Methods

Quantum mechanical calculations

To explain the working mechanism of the electric molecular motor, quantum mechanical calculations were performed at the level of density functional theory to study the PESs of the [2]catenanes **[2]C$^{9+/5+4•}$** and the [3]catenanes **[3]CMM$^{13+/7+6•}$**. In these calculations, the geometries were optimized in the Poisson–Boltzmann solvation model[48] at the level of the M06-2X/6-31G* basis set (ref. [49]) with Jaguar v.10.6 (ref. [50]). Because of the complexity of the systems, all counterions were replaced by an implicit continuous dielectric solvent. To compensate for the effect of counterions, parameters resulting in a stronger solvation effect were necessary. The new solvation parameters were obtained (Supplementary Fig. 36) by fitting the experimentally measured barrier height of the CBPQT^{4+} dethreading from a model pseudorotaxane. The solvation parameters $\varepsilon = 75$ and $R_0 = 1.4$ Å were chosen for the calculations performed on **[2]C$^{9+/5+4•}$** and **[3]CMM$^{13+/7+6•}$**.

PESs of [2]catenane

The PESs of **[2]C$^{9+/5+4•}$** were studied (Extended Data Figs. 1 and 2) by scanning the *z* coordinate from the centre of CBPQT$^{4+/2(+•)}$, which is defined as the average position of four methylene carbon atoms of CBPQT$^{4+/2(+•)}$, passing over the atoms—labelled as positions 0 to 50—on the loop. The PESs are periodic because of the cyclic nature of the loop.

For the oxidized state **[2]C^{9+}**, the PES (Extended Data Fig. 1 and Supplementary Table 1) reaches a maximum (position 31) between the positively charged V^{2+} and PY^{+} units, indicating strong electrostatic repulsion between the CBPQT^{4+} ring and the V^{2+}/PY^{+} units on the loop. The other barrier (position 2) is provided by the IPP unit because of the bulky isopropyl group. Whereas the PES shows a minimum (position 0) positioned around the T unit, the other energy well (position 18) is close to the centre of the BPM unit because of donor–acceptor and van der Waals interactions between the CBPQT^{4+} ring and the BPM unit. The energy barriers for the CBPQT^{4+} ring encircling the BPM unit passing over the bulky IPP and the positively charged PY^{+} units are 25.5 and 46.7 kcal mol^{-1}, respectively.

For the reduced state **[2]C$^{5+4•}$**, the PES (Extended Data Fig. 2a) reaches a maximum (42.2 kcal mol^{-1}) when the CBPQT$^{2(+•)}$ ring passes (position 2) over the bulky IPP unit. All three wells (Extended Data Fig. 2a) result from favourable radical-pairing interactions between the CBPQT$^{2(+•)}$ ring and the V$^{+•}$ units: the first (position 10) and the third (position 27) wells correspond to the CBPQT$^{2(+•)}$ ring encircling the two V$^{+•}$ units (Extended Data Fig. 2d), respectively, whereas the second well (position 21) has a compacted conformation (Extended Data Fig. 2d) because of the radical-pairing interactions between the V$^{+•}$ units and the CBPQT$^{2(+•)}$ ring, which are tilted with respect to each other.

PESs of [3]catenane

To describe the movement of the two CBPQT$^{4+/2(+•)}$ rings in **[3]CMM$^{13+/7+6•}$** around the loop, a two-dimensional map was constructed (Extended Data Fig. 5), in which the *x* and *y* axes represent the positions of the CBPQT$^{4+/2(+•)}$ rings on the loop. Note that the map is periodic in both dimensions. To simplify the calculations, one of the two CBPQT$^{4+/2(+•)}$ rings

(ring A or ring B) was moved to its next position, and then the second ring was allowed to relax to its local minimum. The PESs of **[3]CMM$^{13+/7+6•}$** on the two-dimensional map were calculated by scanning the *z* coordinate from the centre of one CBPQT$^{4+/2(+•)}$ ring passing over the labelled atoms on the loop while letting every other degree of freedom, including the position of the other CBPQT$^{4+/2(+•)}$ ring, relax to the local minimum. Eight hypothetical paths were calculated (Supplementary Tables 3–10) for redox switching in the [3]catenanes **[3]CMM$^{13+/7+6•}$**, which are identified in Extended Data Figs. 3b and 4b. In the calculations, it is assumed that all of the reductions and oxidations occur rapidly relative to the ring motion.

In the case of the reduction process (Extended Data Fig. 3), although all four paths experience a decrease in energy at the beginning, path R3 requires passage over the lowest barrier. In path R3, the CBPQT$^{2(+•)}$ ring B was moved first of all to the bottom V$^{+•}$ unit, followed by moving the CBPQT$^{2(+•)}$ ring A to pass (Extended Data Fig. 3e) over the PY^{+} unit to reach the end point II. In comparison, both paths R2 and R4 require the CBPQT$^{2(+•)}$ ring A to pass (Extended Data Fig. 3f) over the bulky IPP unit from point I to II′. As a result, the path-determining energy difference along path R4 is 4.7 kcal mol^{-1} higher than that along path R3. These results indicate that, under reducing conditions, the CBPQT$^{2(+•)}$ rings (1 and 2) strongly prefer to move from point I to II.

The X-ray single-crystal structure (Fig. 2b) of the reduced [3]catenane **[3]CMM$^{7+6•}$** clearly shows that two CBPQT$^{2(+•)}$ rings encircle the two V$^{+•}$ units in the loop. To study the unidirectional movement of the CBPQT^{4+} rings under oxidizing condition, point II was set to be the starting point. The single-crystal structure was also used as the initial structure for optimizing the geometry of the [3]catenane **[3]CMM^{13+}**. On oxidation, the two CBPQT^{4+} rings move away from the V^{2+} units. The question is, which direction is the more favourable one?

Four paths that all start at point II in two different directions were examined (Extended Data Fig. 4). Among the four paths, path O1 has the lowest energy barrier for the CBPQT^{4+} rings moving towards the end point I′, in which two CBPQT^{4+} rings encircle the T and BPM units. In path O1, the

$CBPQT^{4+}$ ring B was first moved to the BPM unit, followed by moving the $CBPQT^{4+}$ ring A to the T unit, so as to lower the overall energy. By comparison, in path O2, the $CBPQT^{4+}$ ring A was first moved to pass over the PY^{+} unit to its final location at the T unit, followed by moving the $CBPQT^{4+}$ ring A to the BPM unit. The path-determining energy difference along path O2 is 8.1 kcal mol^{-1} higher than that along path O1. Therefore, it is far more favourable for the two $CBPQT^{4+}$ rings to move from point II to point I′.

In conclusion, the quantum mechanical calculations predict unidirectional movement from point I to point II to point I′ during the redox cycle of the [3]catenanes **[3]CMM**$^{\mathbf{13+/7+6\bullet}}$, which is consistent with the experimental result.

Cyclic voltammetry

To gain a better understanding of the electron-transfer processes during the redox cycle undergone by the [3]catenane **[3]CMM**$^{\mathbf{13+}}$, variable scan-rate cyclic voltammetry experiments (Extended Data Fig. 6) were performed.

The cyclic voltammetry profile shows (Extended Data Fig. 6) three reduction peaks with potentials at −0.08 V, −0.15 V and −0.25 V at a low scan rate (0.02 V s^{-1}), corresponding to reduction associated with radical formation, starting from **[3]CMM**$^{\mathbf{13+}}$ and leading to the production of **[3]CMM**$^{\mathbf{7+6\bullet}}$. The first two reduction peaks (−0.08 V and −0.15 V) account for the stepwise formation of viologen radical pairs as a consequence of the different chemical environments experienced by the two $CBPQT^{4+}$ rings in the [3]catenane **[3]CMM**$^{\mathbf{13+}}$. The first reduction peak at −0.08 V can be assigned to the reduction of one of the V^{2+} units in the loop and one of the two V^{2+} units in the $CBPQT^{4+}$ ring encircling the BPM unit, resulting in a decrease in Coulombic repulsion, while establishing stabilizing radical-pairing interactions between the mechanically interlocked components. The following reduction peak at −0.15 V can be attributed to the reduction of the other V^{2+} unit in the loop and one of the two V^{2+} units in the other $CBPQT^{4+}$ ring encircling the T unit. The third reduction peak observed at −0.25 V, corresponding to two simultaneous one-electron reductions,

accounts for the further reduction of both $CBPQT^{2+(\bullet+)}$ monoradical trication rings to their diradical dicationic states $CBPQT^{2(\bullet+)}$. The oxidation of the radical state **[3]CMM$^{7+6\bullet}$** back to **[3]CMM^{13+}** occurs in two steps at −0.16 V and +0.08 V. The first oxidation peak at −0.16 V can be assigned to two one-electron oxidations of the two unpaired $V^{+\bullet}$ units in the two $CBPQT^{(2+\bullet)}$ rings ($CBPQT^{2(+\bullet)} \rightarrow CBPQT^{2+(+\bullet)}$), resulting in much weaker binding interactions and increased Coulombic repulsion between the components ($V^{+\bullet}$ and $CBPQT^{2+(+\bullet)}$), that is, **[3]CMM$^{7+6\bullet}$** is oxidized to **[3]CMM$^{5+4(+\bullet)}$**. It is followed by four simultaneous one-electron oxidations, namely two $V^{+\bullet}$ units in the two $CBPQT^{2+(\bullet+)}$ rings and two $V^{\bullet+}$ units in the loop are oxidized at the same oxidation potential (+0.08 V), resulting in the formation of the fully oxidized state **[3]CMM^{13+}**. These observations are consistent with previously published results[51].

The reduction of the radical state **[3]CMM$^{7+6\bullet}$** to its neutral (viologen) form **[3]CMM^{+}** involves (Supplementary Fig. 46) three sequential two-electron reversible processes. The first and less negative one (−0.77 V, peak potential) can be assigned to the reduction of the two unpaired $V^{+\bullet}$ units in the two $CBPQT^{2(+\bullet)}$ rings, leading to the formation of **[3]CMM$^{+4(+\bullet)}$**, which is stabilized[52] by both radical-pairing and donor–acceptor interactions. The following two-electron process accounts for the reduction of the other two $V^{+\bullet}$ units in the two CBPQT $^{(+\bullet)}$ rings. Finally, the reduction of the remaining two $V^{+\bullet}$ units in the loop occurs at −1.01 V, reflecting the presence of mechanical bonding and the nanoconfined geometry of the [3]catenane.

As the scan rate is increased to 2.0 V s^{-1}, only one reduction peak for the radical state **[3]CMM$^{7+6\bullet}$** is observed, indicating that the electron-transfer process is much faster than ring movement at this fast scan rate. This observation suggests that, under the experimental conditions used during the operation of the [3]catenane motor, the reduction to the radical state **[3]CMM$^{7+6\bullet}$** and the reoxidation to the fully oxidized state **[3]CMM^{13+}** is completed fully and very rapidly.

Electrically driven operation of [3]CMM

Electrically driven operation of the [3]catenane molecular motor **[3]CMM** was conducted in the N_2-filled glovebox. An MeCN (38 ml, 0.1 M $TBAPF_6$) solution of **[3]CMM**•$13PF_6$ (30 μM) was added to a BASi bulk electrolysis cell, which was equipped (Extended Data Fig. 7) with a reticular vitreous carbon working electrode, a coiled platinum-wire auxiliary electrode within a fritted glass chamber and a Ag/AgCl reference electrode, and connected to a Gamry multipurpose instrument (Reference 600) interfaced to a PC. The auxiliary electrode chamber was filled with an excess of $Cu(MeCN)_4PF_6$ dissolved in MeCN (1 ml, 0.1 M $TBAPF_6$). The auxiliary electrode constituted a platinum wire wrapped with a copper wire (diameter 0.25 mm, 99.999% trace metals basis from Sigma-Aldrich). The experimental parameters were controlled using the software of Gamry Framework v.6.30 operating in the chronocoulometry mode. A less negative reduction potential (−0.5 V) and a less positive oxidation potential (+0.7 V) were used to limit the degradation of the [3]catenane. The whole apparatus was subjected to five redox cycles with alternate constant potentials of −0.5 V (reduction potential versus Ag/AgCl) and +0.7 V (oxidation potential versus Ag/AgCl) for 10 and 15 min, respectively.

Data availability

The data that support the findings of this study are available within the paper and its Supplementary Information files. Crystallographic data for the [3]catenane in its reduced state **[3]CMM**$^{7+6•}$ can be obtained free of charge from www.ccdc.cam.ac.uk under CCDC deposition number 2168726.

References

1. Astumian, R. D. Making molecules into motors. *Sci. Am.* **285**, 56–64 (2001).

2. Kassem, S. et al. Artificial molecular motors. *Chem. Soc. Rev.* **46**, 2592–2621 (2017).

3. Baroncini, M., Silvi, S. & Credi, A. Photo- and redox-driven artificial molecular motors. *Chem. Rev.* **120**, 200–268 (2020).

4. Leigh, D. A., Wong, J. K. Y., Dehez, F. & Zerbetto, F. Unidirectional rotation in a mechanically interlocked molecular rotor. *Nature* **424**, 174–179 (2003).

5. Erbas-Cakmak, S. et al. Rotary and linear molecular motors driven by pulses of a chemical fuel. *Science* **358**, 340–343 (2017).

6. Odell, B. et al. Cyclobis(paraquat-*p*-phenylene). A tetracationic multipurpose receptor. *Angew. Chem. Int. Edn Engl.* **27**, 1547–1550 (1988).

7. Astumian, R. D. & Derényi, I. Fluctuation driven transport and models of molecular motors and pumps. *Eur. Biophys. J.* **27**, 474–489 (1998).

8. Chatterjee, M. N., Kay, E. R. & Leigh, D. A. Beyond switches: ratcheting a particle energetically uphill with a compartmentalized molecular machine. *J. Am. Chem. Soc.* **128**, 4058–4073 (2006).

9. Mukherjee, S. & Warshel, A. The F_OF_1 ATP synthase: from atomistic three-dimensional structure to the rotary-chemical function. *Photosynth. Res.* **134**, 1–15 (2017).

10. Astumian, R. D. & Robertson, B. Imposed oscillations of kinetic barriers can cause an enzyme to drive a chemical reaction away from equilibrium. *J. Am. Chem. Soc.* **115**, 11063–11068 (1993).

11. Pezzato, C. et al. Controlling dual molecular pumps electrochemically. *Angew. Chem. Int. Edn* **57**, 9325–9329 (2018).

12. Pumm, A.-K. et al. A DNA origami rotary ratchet motor. *Nature* **607**, 492–498 (2022).

13. Cheng, C., McGonigal, P. R., Stoddart, J. F. & Astumian, R. D. Design and synthesis of nonequilibrium systems. *ACS Nano* **9**, 8672–8688 (2015).

14. Feng, L. et al. Active mechanisorption driven by pumping cassettes. *Science* **374**, 1215–1221 (2021).

15. Balzani, V., Credi, A., Raymo, F. M. & Stoddart, J. F. Artificial molecular machines. *Angew. Chem. Int. Edn* **39**, 3348–3391 (2000).

16. Kay, E. R. & Leigh, D. A. Rise of the molecular machines. *Angew. Chem. Int. Edn* **54**, 10080–10088 (2015).

17. Feringa, B. L. The art of building small: from molecular switches to motors (Nobel Lecture). *Angew. Chem. Int. Edn* **56**, 11060–11078 (2017).

18. Sauvage, J.-P. From chemical topology to molecular machines (Nobel Lecture). *Angew. Chem. Int. Edn* **56**, 11080–11093 (2017).

19. Stoddart, J. F. Mechanically interlocked molecules (MIMs)—molecular shuttles, switches, and machines (Nobel Lecture). *Angew. Chem. Int. Edn* **56**, 11094–11125 (2017).

20. Abendroth, J. M., Bushuyev, O. S., Weiss, P. S. & Barrett, C. J. Controlling motion at the nanoscale: rise of the molecular machines. *ACS Nano* **9**, 7746–7768 (2015).

21. Astumian, R. D. Stochastically pumped adaptation and directional motion of molecular machines. *Proc. Natl. Acad. Sci. USA* **115**, 9405–9413 (2018).

22. García-López, V., Liu, D. & Tour, J. M. Light-activated organic molecular motors and their applications. *Chem. Rev.* **120**, 79–124 (2020).

23. Dattler, D. et al. Design of collective motions from synthetic molecular switches, rotors, and motors. *Chem. Rev.* **120**, 310–433 (2020).

24. Ragazzon, G., Baroncini, M., Silvi, S., Venturi, M. & Credi, A. Light-powered autonomous and directional molecular motion of a dissipative self-assembling system. *Nat. Nanotechnol.* **10**, 70–75 (2015).

25. Cheng, C. et al. An artificial molecular pump. *Nat. Nanotechnol.* **10**, 547–553 (2015).

26. Qiu, Y. et al. A precise polyrotaxane synthesizer. *Science* **368**, 1247–1253 (2020).

27. Amano, S., Fielden, S. D. P. & Leigh, D. A. A catalysis-driven artificial molecular pump. *Nature* **594**, 529–534 (2021).

28. Hernández, J. V., Kay, E. R. & Leigh, D. A. A reversible synthetic rotary molecular motor. *Science* **306**, 1532–1537 (2004).

29. Wilson, M. R. et al. An autonomous chemically fuelled small-molecule motor. *Nature* **534**, 235–240 (2016).

30. Pooler, D. R. S., Lubbe, A. S., Crespi, S. & Feringa, B. L. Designing light-driven rotary molecular motors. *Chem. Sci.* **12**, 14964–14986 (2021).

31. Koumura, N., Zijlstra, R. W. J., van Delden, R. A., Harada, N. & Feringa, B. L. Light-driven monodirectional molecular rotor. *Nature* **401**, 152–155 (1999).

32. Štacko, P. et al. Locked synchronous rotor motion in a molecular motor. *Science* **356**, 964–968 (2017).

33. Amano, S., Borsley, S., Leigh, D. A. & Sun, Z. Chemical engines: driving systems away from equilibrium through catalyst reaction cycles. *Nat. Nanotechnol.* **16**, 1057–1067 (2021).

34. Borsley, S., Kreidt, E., Leigh, D. A. & Roberts, B. M. W. Autonomous fuelled directional rotation about a covalent single bond. *Nature* **604**, 80–85 (2022).

35. Kudernac, T. et al. Electrically driven directional motion of a four-wheeled molecule on a metal surface. *Nature* **479**, 208–211 (2011).

36. Tierney, H. L. et al. Experimental demonstration of a single-molecule electric motor. *Nat. Nanotechnol.* **6**, 625–629 (2011).

37. Perera, U. G. E. et al. Controlled clockwise and anticlockwise rotational switching of a molecular motor. *Nat. Nanotechnol.* **8**, 46–51

(2012).

38. Trabolsi, A. et al. Radically enhanced molecular recognition. *Nat. Chem.* **2**, 42–49 (2010).

39. Cai, K., Zhang, L., Astumian, R. D. & Stoddart, J. F. Radical-pairing-induced molecular assembly and motion. *Nat. Rev. Chem.* **5**, 447–465 (2021).

40. Pezzato, C. et al. An efficient artificial molecular pump. *Tetrahedron* **73**, 4849–4857 (2017).

41. Qiu, Y. et al. A molecular dual pump. *J. Am. Chem. Soc.* **141**, 17472–17476 (2019).

42. Qiu, Y., Feng, Y., Guo, Q.-H., Astumian, R. D. & Stoddart, J. F. Pumps through the ages. *Chem* **6**, 1952–1977 (2020).

43. Feng, Y. et al. Molecular pumps and motors. *J. Am. Chem. Soc.* **143**, 5569–5591 (2021).

44. Bruns, C. J. & Stoddart, J. F. *The Nature of the Mechanical Bond: From Molecules to Machines* (Wiley, 2016).

45. Astumian, R. D. & Bier, M. Biased Brownian motion as the operating principle for microscopic engines. *Bioelectrochem. Bioenerg.* **39**, 67–75 (1996).

46. Grommet, A. B., Feller, M. & Klajn, R. Chemical reactivity under nanoconfinement. *Nat. Nanotechnol.* **15**, 256–271 (2020).

47. Astumian, R. D. Thermodynamics and kinetics of a Brownian motor. *Science* **276**, 917–922 (1997).

48. Tannor, D. J. et al. Accurate first principles calculation of molecular charge distributions and solvation energies from ab initio quantum mechanics and continuum dielectric theory. *J. Am. Chem. Soc.* **116**, 11875–11882 (1994).

49. Zhao, Y. & Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **120**, 215–241 (2007).

50. Jaguar v.10.6 (Schrödinger, Inc., 2019).

51. Li, H. et al. Mechanical bond-induced radical stabilization. *J. Am. Chem. Soc.* **135**, 456–467 (2013).

52. Jiao, Y. et al. Electron-catalysed molecular recognition. *Nature* **603**, 265–270 (2022).

Acknowledgements

We thank Northwestern University (NU) for its support of this research, the Integrated Molecular Structure Education and Research Center (IMSERC) at NU for providing access to equipment for relevant experiments, S. Shafaie for assistance with high-resolution mass spectrometry measurements, Y. Zhang and Y. Wu for help with NMR spectroscopic measurements and C. Cheng for useful discussions. The computational investigations at California Institute of Technology were supported by National Science Foundation grant no. CBET-2005250 (W.-G.L. and W.A.G.).

Author information

Authors and Affiliations

1. Department of Chemistry, Northwestern University, Evanston, IL, USA

   Long Zhang, Yunyan Qiu, Hongliang Chen, Dengke Shen, Bo Song, Kang Cai, Huang Wu, Yang Jiao, Yuanning Feng, James S. W. Seale, Cristian Pezzato, Yu Tan, Xiao-Yang Chen, Qing-Hui Guo, Charlotte L. Stern, Douglas Philp & J. Fraser Stoddart

2. Materials and Process Simulation Center, California Institute of Technology, Pasadena, CA, USA

Wei-Guang Liu & William A. Goddard III

3. Stoddart Institute of Molecular Science, Department of Chemistry, Zhejiang University, Hangzhou, China

Hongliang Chen, Qing-Hui Guo & J. Fraser Stoddart

4. ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China

Hongliang Chen, Qing-Hui Guo & J. Fraser Stoddart

5. Institutes of Physical Science and Information Technology, Anhui University, Hefei, China

Dengke Shen

6. Department of Chemistry, Nankai University, Tianjin, China

Kang Cai

7. Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Cristian Pezzato

8. Department of Chemical Sciences, University of Padova, Padova, Italy

Cristian Pezzato

9. Key Laboratory of Synthetic and Self-Assembly Chemistry for Organic Functional Molecules, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, China

Jia Tian

10. School of Chemical Engineering and Technology, Sun Yat-sen University, Zhuhai, China

Yu Tan

11. School of Chemistry, University of St Andrews, North Haugh, St Andrews, UK

Douglas Philp

12. Department of Physics and Astronomy, University of Maine, Orono, ME, USA

R. Dean Astumian

13. School of Chemistry, University of New South Wales, Sydney, New South Wales, Australia

J. Fraser Stoddart

Contributions

J.F.S. directed the project. L.Z. and C.P. conceived the idea for the project. L.Z. designed, synthesized and characterized the compounds. W.-G.L. and W.A.G. performed density functional theory calculations. R.D.A. contributed to theoretical analyses carried out on the mechanism of operation of the electric molecular motor. Y.Q. provided some of the precursors and performed CPE experiments. H.C. contributed to the graphical design used in the figures. D.S. conducted the video recording and contributed to the animation. B.S. carried out electrospray ionization mass spectrometry and travelling-wave ion mobility mass spectrometry measurements and analyses. C.L.S. collected the single-crystal X-ray diffraction data and solved the solid-state structure. K.C., H.W., Y.J., Y.F., J.S.W.S., J.T., Y.T., X.-Y.C., Q.-H.G. and D.P. commented on the data. All the authors participated in evaluating the results. L.Z., R.D.A. and J.F.S. produced numerous drafts of the manuscript and supplementary materials, with input from all authors.

Corresponding authors

Correspondence to Long Zhang, R. Dean Astumian, William A. Goddard III or J. Fraser Stoddart.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Junpei Yuasa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Calculated PES of the oxidized [2]catenane [2]C^{9+}.

a, The potential energy for the CBPQT^{4+} ring traversing the loop in the oxidized state. The numbered atoms on the loop are used to define the position of the CBPQT^{4+} ring. **b**, Graphical representation of the oxidized state of the [2]catenane **[2]C^{9+}**. **c**, Graphical representation of the calculated

PES of the $CBPQT^{4+}$ ring moving around the loop, shown in a rollercoaster manner for the fully oxidized **[2]C**$^{9+}$.

Extended Data Fig. 2 Calculated PES of the reduced [2]catenane [2]C$^{5+4•}$.

a, The potential energy for the $CBPQT^{2(+•)}$ ring as a function of its position on the loop in the reduced state of the [2]catenane **[2]C**$^{\mathbf{5+4•}}$. The numbered atoms on the loop are used to define the position of the $CBPQT^{2(+•)}$ ring. **b**, Graphical representation of the reduced state of [2]catenane **[2]C**$^{\mathbf{5+4•}}$ in its lowest-energy co-conformation in which the $CBPQT^{2(+•)}$ ring encircles the $V^{+•}$ unit, position 10. **c**, Graphical representation of calculated potential energy of the $CBPQT^{\mathbf{2(+•)}}$ ring as it moves around the loop, shown in a rollercoaster manner for the radical state **[2]C**$^{\mathbf{5+4•}}$. **d**, Quantum mechanical minimized structures (M06-2X/6-31G* basis set) for the $CBPQT^{2(+•)}$ ring located at positions 10, 21 and 27, respectively.

Extended Data Fig. 3 Calculated PES of the reduced [3]catenane [3]CMM$^{7+6•}$.

a, Graphical representations of the redox process experienced by the [3]catenane in going from the oxidized state I (48, 19) to the reduced state II (28, 10). **b**, A two-dimensional position map describing the movement of the two reduced $CBPQT^{2(+•)}$ rings (A and B) around the loop. Four hypothetical paths (R1–R4) during the reduction process are illustrated by diamond (grey), circle (light red), triangle (black) and square (red) symbols, respectively. The green arrow indicates the preferred direction of movement and the red arrow indicates the less preferred (nearly precluded) direction of movement. The green, purple, black, blue and magenta lines represent the positions of the IPP, $V^{+•}$, BPM, PY^{+} and T units, respectively. The dashed red diagonal lines represent barriers that cannot be crossed physically because doing so would require the two $CBPQT^{2(+•)}$ rings to occupy the same space. **c**, The PESs of the two $CBPQT^{2(+•)}$ rings moving around the loop in the reduced state, starting from point I (48, 19) and following paths R3 and R4, respectively. **d**, Structural formula of the loop with atoms

numbered to define the positions of the $CBPQT^{2(+\bullet)}$ rings. **e**,**f**, Quantum mechanical minimized structures (M06-2X/6-31G* basis set, top-down and side-on views) for the $CBPQT^{2(+\bullet)}$ rings at positions (32, 10) and (2, 28), respectively.

Extended Data Fig. 4 Calculated PES of the oxidized [3]catenane [3]CMM^{13+}.

a, Graphical representations of the process experienced by the [3]catenane in going from the oxidized state II (28, 10) to the oxidized state I′ (19, 49). **b**, A two-dimensional position map describing the movement of the two oxidized $CBPQT^{4+}$ rings (A and B) around the loop. Four hypothetical paths O1–O4 during the oxidation process are illustrated by diamond (black), circle (red), triangle (grey) and square (light red) symbols, respectively. The green arrow indicates the preferred direction of movement and the red arrow indicates the less preferred (nearly precluded) direction of movement. The green, blue, black, blue and magenta lines represent the positions of the IPP, V^{2+}, BPM, PY^{+} and T units, respectively. The dashed red diagonal line represents a barrier that cannot be crossed physically because the two $CBPQT^{4+}$ rings would occupy the same space. **c**, Structural formula of the loop with atoms numbered to define the positions of the $CBPQT^{4+}$ rings. **d**, The PESs of the two $CBPQT^{4+}$ rings moving around the loop in the oxidized state starting from point II (28, 10) and following paths O1 and O2, respectively. The value for the energy barrier $\Delta E^{\ddagger}$ of 20.1 kcal mol^{-1} was determined from the energy difference between the positions (19, 3) and (20, 1). The position (19, 3) corresponds to the metastable state on path O1. **e**,**f**, Quantum mechanical minimized structures (M06-2X/6-31G* basis set, top-down and side-on views) for the $CBPQT^{4+}$ rings at positions (20, 1) and (34, 10), respectively. Quantum mechanical minimized structures of the lowest-energy state (19, 48) and metastable state (21, 3) are presented in Figs. 2a and 4a, respectively.

Extended Data Fig. 5 A two-dimensional position map describing the movement of two $CBPQT^{4+/2(+\bullet)}$ rings around the loop.

The *x* and *y* axes represent the positions of the $CBPQT^{4+/2(+•)}$ rings on the loop. The green, black, blue and magenta lines represent the positions of the IPP, BPM, PY^{+} and T units, respectively. The hypothetical paths involving switching of the [3]catenane **[3]CMM** during the redox cycles are illustrated in blue (oxidation) with purple (reduction) arrows, which are periodic on account of the cyclic nature of the loop. The trajectories indicate that, for unidirectional motion, there is a Coulombic barrier (PY^{+}) and a steric barrier (IPP) under reducing and oxidizing conditions, respectively. The dashed red diagonal line represents a barrier that cannot be crossed physically for the simple reason that two $CBPQT^{4+/2(+•)}$ rings would end up occupying the same space.

Extended Data Fig. 6 Scan-rate variation (0.02–2.0 V s^{-1}) of cyclic voltammograms of [3]CMM•13PF_6 (0.5 mM).

As the scan rate is increased to 2.0 V s^{-1}, only one reduction peak for the radical state **[3]CMM$^{7+6•}$** is observed, indicating that the electron-transfer process is much faster than ring movement at this fast scan rate. This observation suggests that, under the experimental conditions used during the operation of the [3]catenane motor, the reduction to the radical state **[3]CMM$^{7+6•}$** and the reoxidation to the fully oxidized state **[3]CMM^{13+}** is completed fully and very rapidly.

Extended Data Fig. 7 Electrically driven operation of [3]CMM.

a, Graphical illustration of the electrochemical cell used in the repeated controlled potential electrolysis experiments. CE, counter electrode; RE, reference electrode; WE, working electrode. **b**, Photographs of the oxidized state (left) **[3]CMM^{13+}** and the reduced state (right) **[3]CMM$^{7+6•}$** in the electrochemical cell.

Extended Data Fig. 8 Synthesis of the deuterium-labelled [3]catenane [D_n]-[3]CMM.

By sequentially adding [D_{16}]-**CBPQT$^{2(+•)}$** and **CBPQT$^{2(+•)}$** during the synthesis of **[D$_n$]-[3]CMM**, it is able to bias the distribution of the undeuterated $CBPQT^{4+}$ rings on the T and BPM units. The [3]catenane [D_n]-**[3]CMM^{13+}** was obtained as a mixture of four isotopologues, including [D_0]-**[3]CMM^{13+}**, and [D_{32}]-**[3]CMM^{13+}**, in addition to two co-constitutional isomers [D_{16}]-**[3]CMM^{13+}**. CuAAC, copper(I)-catalysed azide–alkyne cycloaddition.

Extended Data Fig. 9 Calculation of the probability of 180° rotation for both rings on the loop.

Distribution of $CBPQT^{4+}$ rings on the two positions (T and BPM) of the cyclic track according to the ^{1}H NMR spectra (Fig. 3) of [D_n]-**[3]CMM**•$13PF_6$ before and after one redox cycle, based on the relevant integration of protons H-Phen″ (I_1/I_1') and H-Phen′ (I_2/I_2'), respectively. η, the probability of 180° rotation for both rings on the loop.

Extended Data Fig. 10 Probable direction of the movement of two CBPQT$^{4+/2(+•)}$ rings during the redox operation of [3]CMM^{13+}.

Clockwise means that each CBPQT$^{4+/2(+•)}$ ring sees the substituents on the loop in the order T → PY^+ → $V^{2+/+•}$ → $V^{2+/+•}$ → IPP → T and counterclockwise means that each ring sees the substituents on the loop in the order T → IPP → $V^{2+/+•}$ → $V^{2+/+•}$ → PY^+ → T. The curly arrows represent the direction in which the two rings move around the loop. The 'p_{red}' and 'p_{ox}' near the arrows are the probability of clockwise motion of the two rings on the loop in the redox cycle. The $k_{red,+}$, $k_{red,-}$, $k_{ox,-}$ and $k_{ox,+}$ are the corresponding rate constants for the probable steps during the redox cycle, respectively, and the clockwise and counterclockwise steps are also indicated by '+' and '−', respectively. The 180° rotation is represented by '1/2'.

Supplementary information

Supplementary Information

This file contains Supplementary Figs. 1–78, Schemes 1–9, Tables 1–11, supplementary text and notes, extensive experimental data, quantum mechanical calculation results and detailed discussions.

Supplementary Video 1

Video of controlled potential electrolysis of the [3]catenane molecular motor in an electrochemical cell.

Supplementary Video 2

Animation of the unidirectional rotary motion in the [3]catenane molecular motor during a redox cycle.

Peer Review File

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Zhang, L., Qiu, Y., Liu, WG. *et al.* An electric molecular motor. *Nature* **613**, 280–286 (2023). https://doi.org/10.1038/s41586-022-05421-6

- Received: 26 April 2022
- Accepted: 06 October 2022
- Published: 11 January 2023
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05421-6

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

An artificial molecular motor powered by electricity

Nature Research Briefing 11 Jan 2023

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05421-6

- Article
- Open Access
- Published: 11 January 2023

Seasonal temperatures in West Antarctica during the Holocene

- Tyler R. Jones ,
- Kurt M. Cuffey,
- William H. G. Roberts,
- Bradley R. Markle,
- Eric J. Steig,
- C. Max Stevens,
- Paul J. Valdes,
- T. J. Fudge,
- Michael Sigl,
- Abigail G. Hughes,
- Valerie Morris,
- Bruce H. Vaughn,
- Joshua Garland,
- Bo M. Vinther,
- Kevin S. Rozmiarek,
- Chloe A. Brashear &
- …
- James W. C. White

Show authors

Nature **volume 613**, pages 292–297 (2023)

- 2729 Accesses

- 161 Altmetric

- Metrics details

Abstract

The recovery of long-term climate proxy records with seasonal resolution is rare because of natural smoothing processes, discontinuities and limitations in measurement resolution. Yet insolation forcing, a primary driver of multimillennial-scale climate change, acts through seasonal variations with direct impacts on seasonal climate[1]. Whether the sensitivity of seasonal climate to insolation matches theoretical predictions has not been assessed over long timescales. Here, we analyse a continuous record of water-isotope ratios from the West Antarctic Ice Sheet Divide ice core to reveal summer and winter temperature changes through the last 11,000 years. Summer temperatures in West Antarctica increased through the early-to-mid-Holocene, reached a peak 4,100 years ago and then decreased to the present. Climate model simulations show that these variations primarily reflect changes in maximum summer insolation, confirming the general connection between seasonal insolation and warming and demonstrating the importance of insolation intensity rather than seasonally integrated insolation or season duration[2,3]. Winter temperatures varied less overall, consistent with predictions from insolation forcing, but also fluctuated in the early Holocene, probably owing to changes in meridional heat transport. The magnitudes of summer and winter temperature changes constrain the lowering of the West Antarctic Ice Sheet surface since the early Holocene to less than 162 m and probably less than 58 m, consistent with geological constraints elsewhere in West Antarctica[4,5,6,7].

Main

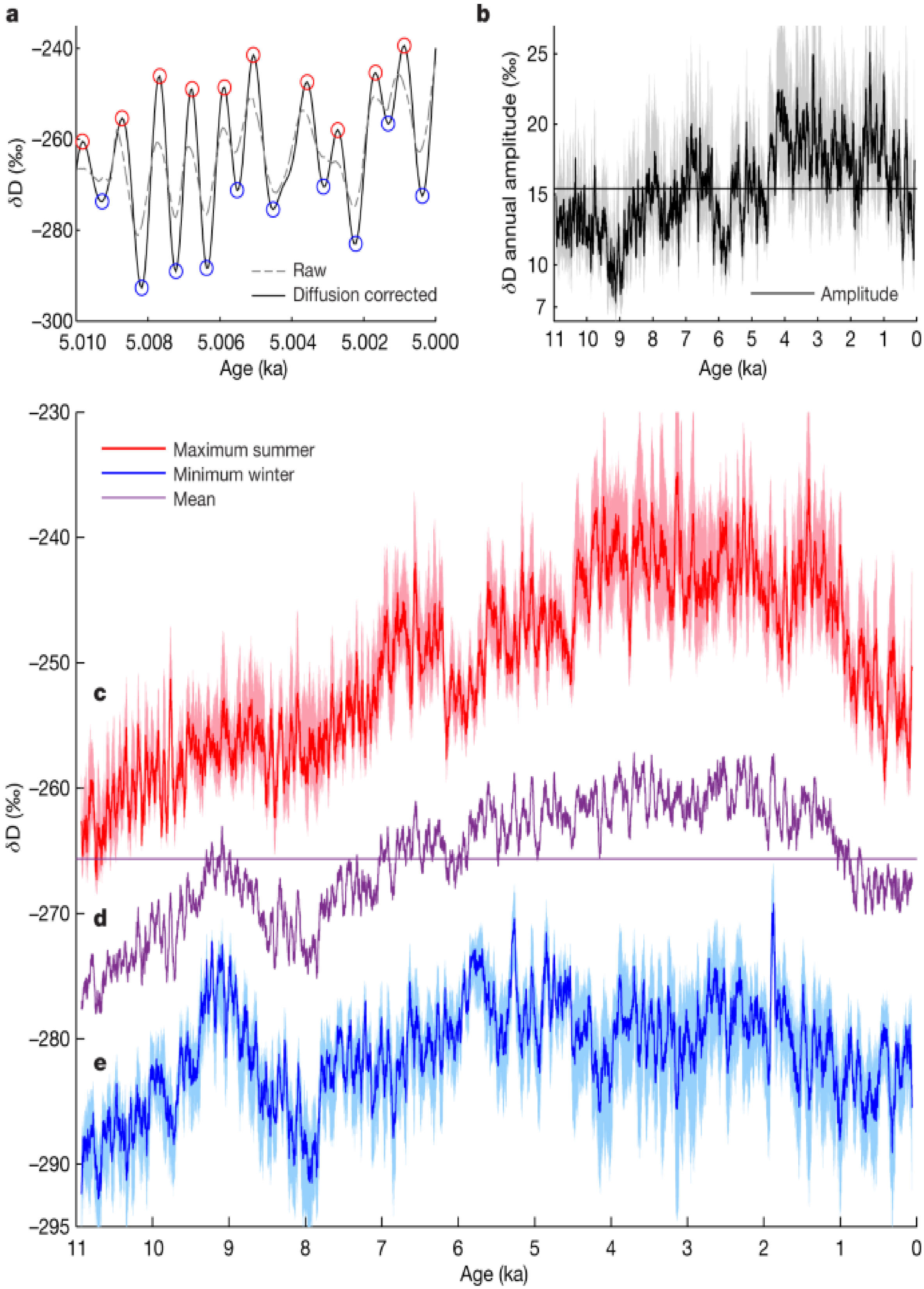
Milankovitch famously postulated that variations of Earth's orbit and axis drive climate changes over tens of thousands of years by altering the seasonal cycle of insolation[1]. By controlling summer temperatures and ice ablation, summer insolation in the northern high latitudes is thought to drive global ice volume changes over glacial–interglacial timescales[8]. Although modelling studies support this idea[9,10], empirical evidence of the specific climate response to insolation changes derives almost entirely from mean

annual temperature reconstructions[11,12] or from indirect effects on, for example, trapped gases and melt layers in polar ice[13,14] and marine aeolian deposits[15]. The absence of seasonal temperature reconstructions has precluded direct evidence of insolation forcing on seasonal climate, a relationship that may vary geographically. In Antarctica, long records of multiple glacial–interglacial cycles have supported different claims about whether the effects of summer insolation relate most strongly to its maximum intensity, its seasonal integral or to duration above a threshold[2,3,16,17]. Site-specific empirical determinations would provide valuable tests of such competing ideas.

Seasonal temperature reconstructions

We reconstructed seasonal temperature variability in West Antarctica through the Holocene (the last 11,000 years) and performed model experiments to understand its physical controls. The Holocene offers a window of time for assessing the influence of orbital forcing without the complicating effects of Northern Hemisphere deglaciation[18]. Our reconstruction (Figs. 1 and 2) uses the high-resolution water-isotope record (δD) from the West Antarctic Ice Sheet (WAIS) Divide ice core (WDC)[18,19,20] (Methods—Water isotopes; Extended Data Fig. 1a,b), obtained with a continuous-flow technique that provides millimetre-scale depth resolution[21]. Layer ages were determined previously[2,22].

Fig. 1: Water-isotope seasonal variability.

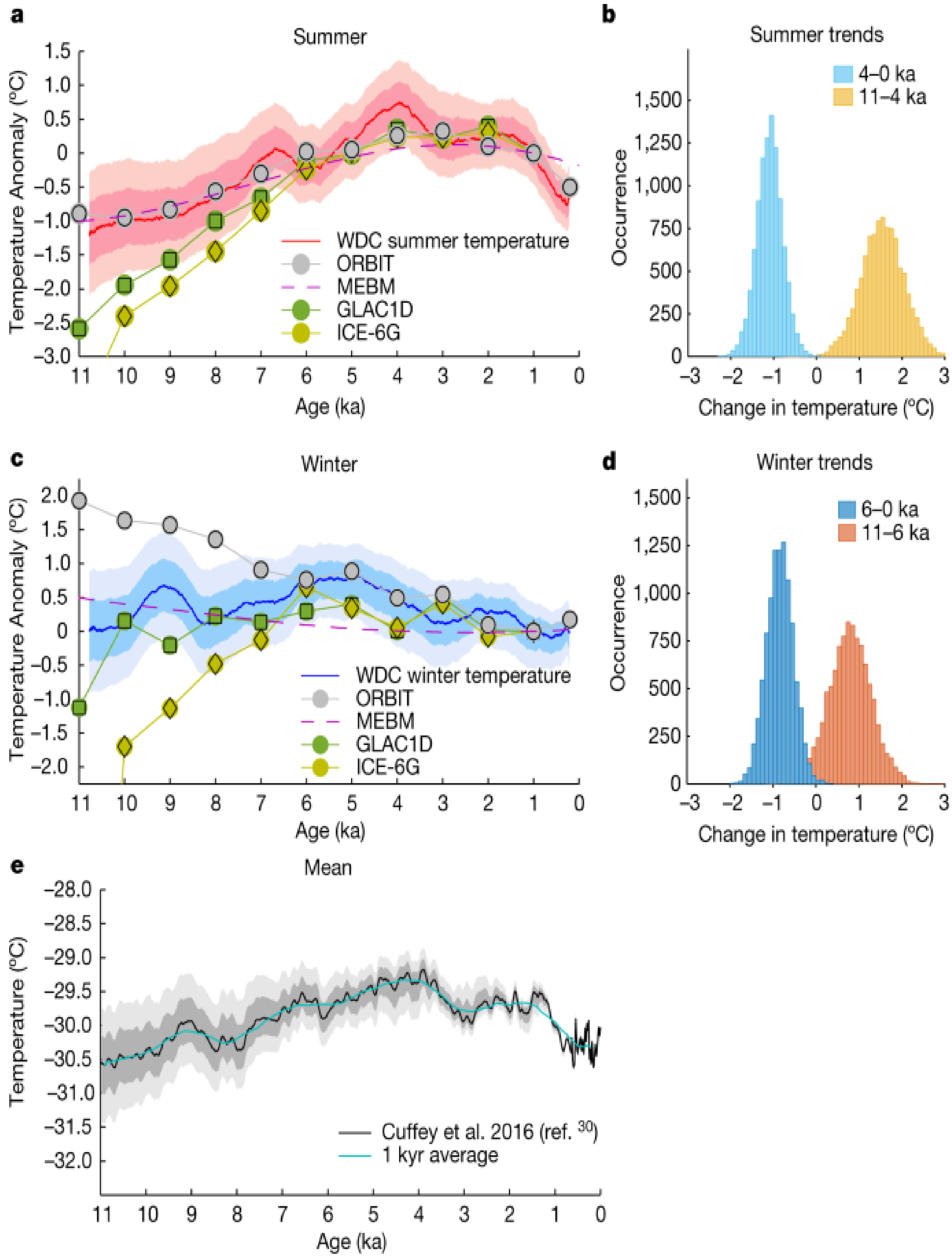


a
b
c
d
e
δD (‰)
Age (ka)
Raw
Diffusion corrected
δD annual amplitude (‰)
Amplitude
Maximum summer
Minimum winter
Mean

a, Example section of the diffusion-corrected (solid line) and raw[20] (dashed line) WDC δD records, with annual maxima (red circles) and minima (blue circles) determined algorithmically (Methods—Seasonal water-isotope amplitudes). Extended Data Fig. 1 provides the full high-resolution WDC δD record, diffusion lengths and extrema. **b**, The 50-yr annual-amplitude averages (summer minus winter divided by 2), with 2σ uncertainty; horizontal line indicates Holocene mean. **c**–**e**, The 50-yr δD averages for summer (red, **c**), mean (purple, **d**) and winter (blue, **e**); horizontal line indicates Holocene mean; shaded regions are 2σ bounds for combined analytical and diffusion-correction uncertainty.

Source data

Fig. 2: Seasonal temperature reconstruction.



a,**c**, Reconstructed summer and winter temperatures at WDC for 1,000-yr averaging (solid red and blue lines). Shaded regions are 1σ and 2σ

uncertainty ranges for combined uncertainties arising from analysis, diffusion correction, seasonality of accumulation, precipitation intermittency, isotope-temperature scaling and reconstructed mean temperatures (Methods—Uncertainties in reconstructing temperatures). Also shown are MEBM-calculated temperatures for 80° S (maximum and minimum annual values) and HadCM3 zonal temperatures for 80° S (late-December for summer, mid-August for winter) (ORBIT, GLAC1D and ICE-6G). The 0 ka ORBIT simulation uses pre-industrial settings, a calculation not available for GLAC1D or ICE-6G. Normalization is done at 1 ka when all model runs intersect within 0.05 °C and the ice-sheet configuration is well known. The ICE-6G values at 11 ka for summer and winter (not shown on plots) are −3.93 °C and −10.82 °C, respectively. Coefficient of determinations for model results versus WDC temperatures (Extended Data Fig. 7) are high for summer (HadCM3 ORBIT $R^2 = 0.93$, $P \ll 0.001$; MEBM $R^2 = 0.80$, $P \ll 0.001$) but not for winter (HadCM3 ORBIT $R^2 = 0.00$, $P = 0.85$; MEBM $R^2 = 0.05$, $P = 0.30$). The winter agreement improves if only the period 0–6 ka is considered (HadCM3 ORBIT $R^2 = 0.74$, $P = 0.01$; MEBM $R^2 = 0.39$, $P = 0.02$). **b**,**d**, Histograms of net temperature changes over the specified time intervals, derived by Monte Carlo analysis accounting for systematic and non-systematic uncertainties (Methods—Trend analysis). **e**, WDC mean annual temperature with 1σ and 2σ uncertainty bounds[30]. Extended Data Table 2 shows the amount of variability in the mean annual temperature that can be explained by the summer and winter temperatures.

Source data

Records of seasonal temperatures from ice cores are limited by measurement resolution and information loss from water-isotope diffusion. In Greenland, the longest records separating summer and winter variability extend to only 2 thousand years ago (ka) (refs. [23,24]),whereas only climate model simulations are available for older periods[10]. For Antarctica, before the present study, the longest records spanned only a few centuries[25]. A combination of three factors accounts for the considerably greater scope of our reconstruction: exceptional depth resolution of measurements, conditions at WAIS Divide (high accumulation, low temperature and thick ice) which allow for preservation of subannual information through the

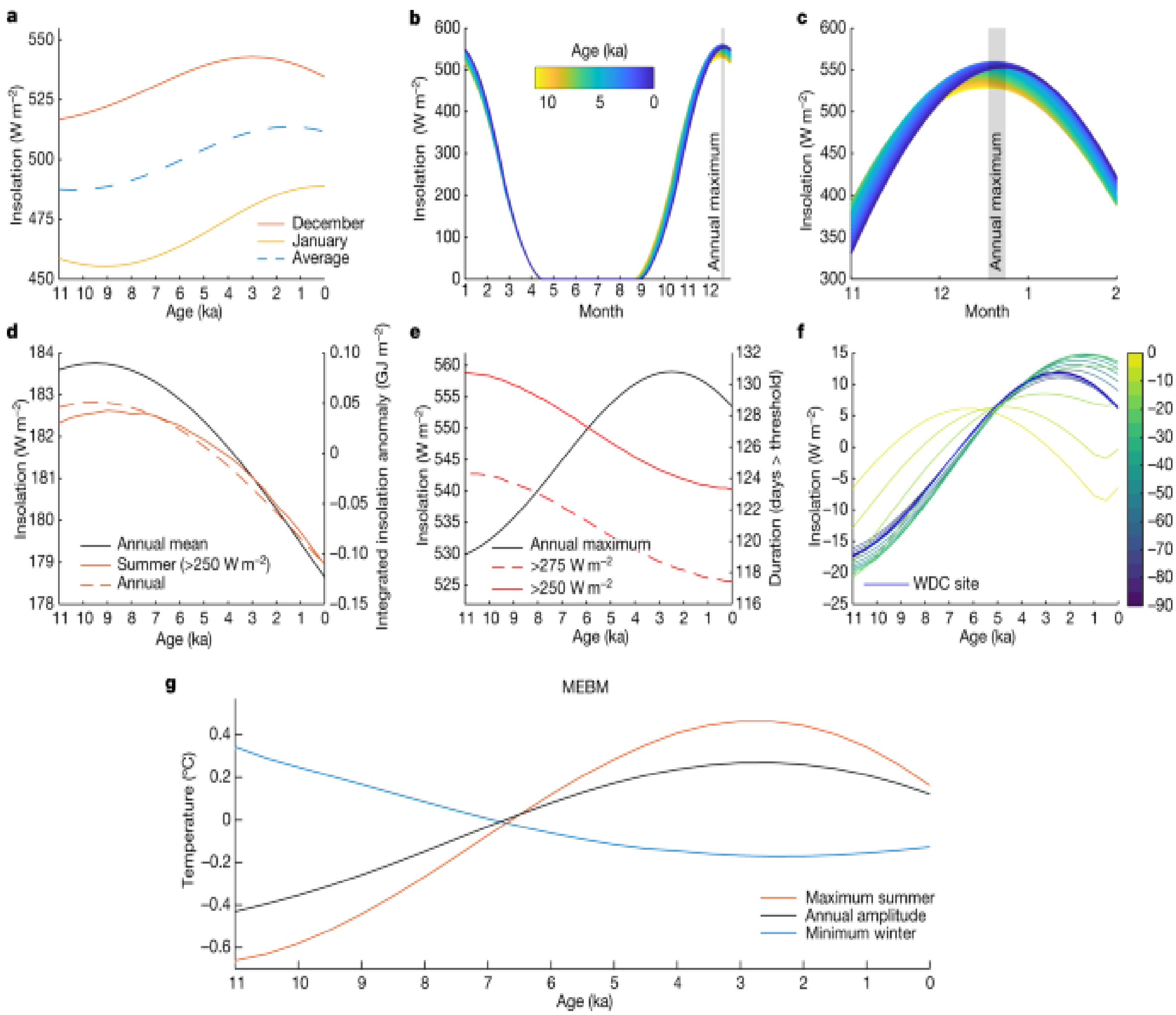
entire Holocene[26] and an analysis strategy which circumvents interannual noise by evaluating millennial averages of the seasonal parameters.

Our method corrects water-isotope variations for diffusion[26,27,28] and assesses uncertainties including preservation bias and precipitation intermittency (Methods—Diffusion corrections and Uncertainties in reconstructing temperatures). The diffusion correction operates on the high-resolution data and produces isotopic time series from which seasonal summer–winter amplitudes were extracted. These were converted to temperature using a model-derived scaling[29] (6.96‰ δD °C^{-1}; Methods—Seasonal temperatures) and added to previously reconstructed annual mean temperatures[30] to obtain summer and winter histories.

Seasonal trends

Summer temperatures at WAIS Divide (Fig. 2a) generally rose through the early and middle Holocene, persisted at a maximum between about 5 and 1.5 ka, then decreased toward the present, with a total Holocene range of around 2 °C. These variations broadly correlate with local maximum insolation, rather than with integrated summer insolation or the duration of summer (Fig. 3d,e). Winter temperatures (Fig. 2c) varied less than summer ones overall (about 1 °C range) but also fluctuated at about 10 to 8 ka, a variation too rapid to attribute to orbital forcing.

Fig. 3: Temporal and spatial variability in insolation and model temperatures.



a, Insolation change through the Holocene[42] for December and January and their average. December best resembles the WDC summer reconstruction. **b**, The full seasonal cycle of insolation at 80° S for 500-yr snapshots over the Holocene. Line colours in **b** and **c** correspond to age. **c**, Zoom of summer insolation. The maximum always occurs in the latter half of December (grey shading), migrating across 8 days over the course of the Holocene. **d**, Holocene trends of annual mean insolation (black), annual integrated insolation (dashed red line) and summer-integrated insolation (red line). **e**, Maximum summer insolation intensity (black line) and summer duration (red lines), defined as the number of days above a threshold insolation value each year. **f**, Anomaly in maximum insolation coloured by latitude in the Southern Hemisphere. The thick blue line shows the latitude of the WDC site. **g**, Calculated temperatures for 80° S using the MEBM, including

maximum summer value (red), minimum winter value (blue) and amplitude of the seasonal temperature cycle (black).

Source data

Annual mean WAIS Divide temperature changes[30] (Fig. 2e) were considerably influenced by winter variability in the early Holocene, whereas summer variability dominates the overall Holocene pattern (Methods—Relationship between the annual mean and individual seasons; Extended Data Table 2). Summer variability also accounts for most of the cooling in the last 2 kyr, indicating that the approximately 1 °C annual-average cooling of the entire West Antarctic during this period[31,32] likewise reflects this season. Neither season at WDC experienced the early Holocene optimum nor overall Holocene cooling that appears in some global temperature reconstructions[33,34]. To assess the significance of the dominant multimillennial trends in each season, we performed Monte Carlo analysis (Methods—Trend analysis) using 4 ka as a demarcation point in summer (this is the timing of maximum summer temperature) and 6 ka in winter (when winter temperatures plateau). For summer (Fig. 2b) this indicates a >95% chance that warming from 11 to 4 ka and cooling from 4 ka to present exceeded 0.7 and 0.6 °C, respectively. For winter, the trend from 11 to 6 ka is indistinguishable from zero, whereas cooling of greater than about 0.3 °C from 6 to 0 ka occurred with >95% likelihood (Fig. 2d).

Moist energy balance model

To evaluate how orbitally driven insolation changes may explain the WAIS Divide reconstructed temperatures (Fig. 2), we first simulated temperature history at 80° S using a global, zonal mean (2° resolution) moist energy balance model (MEBM) accounting for incoming and outgoing radiation, albedo and meridional atmospheric-heat transport (Methods—Moist energy balance model). The model is driven by top-of-atmosphere (TOA) seasonal insolation changes (Fig. 3a–e); for this latitude, the maximum summer insolation increases until about 2.5 ka and annual mean and annual- and summer-integrated values mostly decline through the Holocene. The calculations yield summer maximum temperatures and seasonal temperature amplitudes (Fig. 3g) that covary with local maximum summer insolation

(Fig. 3e) and with the general pattern of our reconstructed summer temperatures (Extended Data Fig. 7). Although heating at lower latitudes can influence Antarctic temperature through atmospheric and oceanic heat transport, modelled maximum summer temperatures at WAIS Divide correlate best with local insolation (70° to 90° S, $R^2 = 0.9$, $P \ll 0.001$ during 0 to 6 ka) rather than insolation anywhere in the subtropical through subpolar latitudes (20° to 60° S, $R^2 = 0.33–0.55$, $P < 0.05$). Indeed, models indicate heat export from WAIS Divide in summer (Extended Data Fig. 4k), rather than import from more northern locations. Since December is always the month of maximum insolation (Fig. 3a–c), variability of December insolation dominates the response of maximum summer temperature. For winter, modelled temperatures are less variable than those of summer at 80° S (Fig. 3g) because of the lack of direct insolation (Fig. 3b) and have an opposite trend. Winter minima are a function of three factors: changes in the length of the zero-insolation season, the effective cooling rate of the surface and convergent heat transport from lower latitudes. Lower minimum winter temperatures occur at times when the zero-insolation season is longer. However, neither the length of the zero-insolation season, modelled minimum temperatures, nor winter heat divergence correlate well with reconstructed winter temperatures.

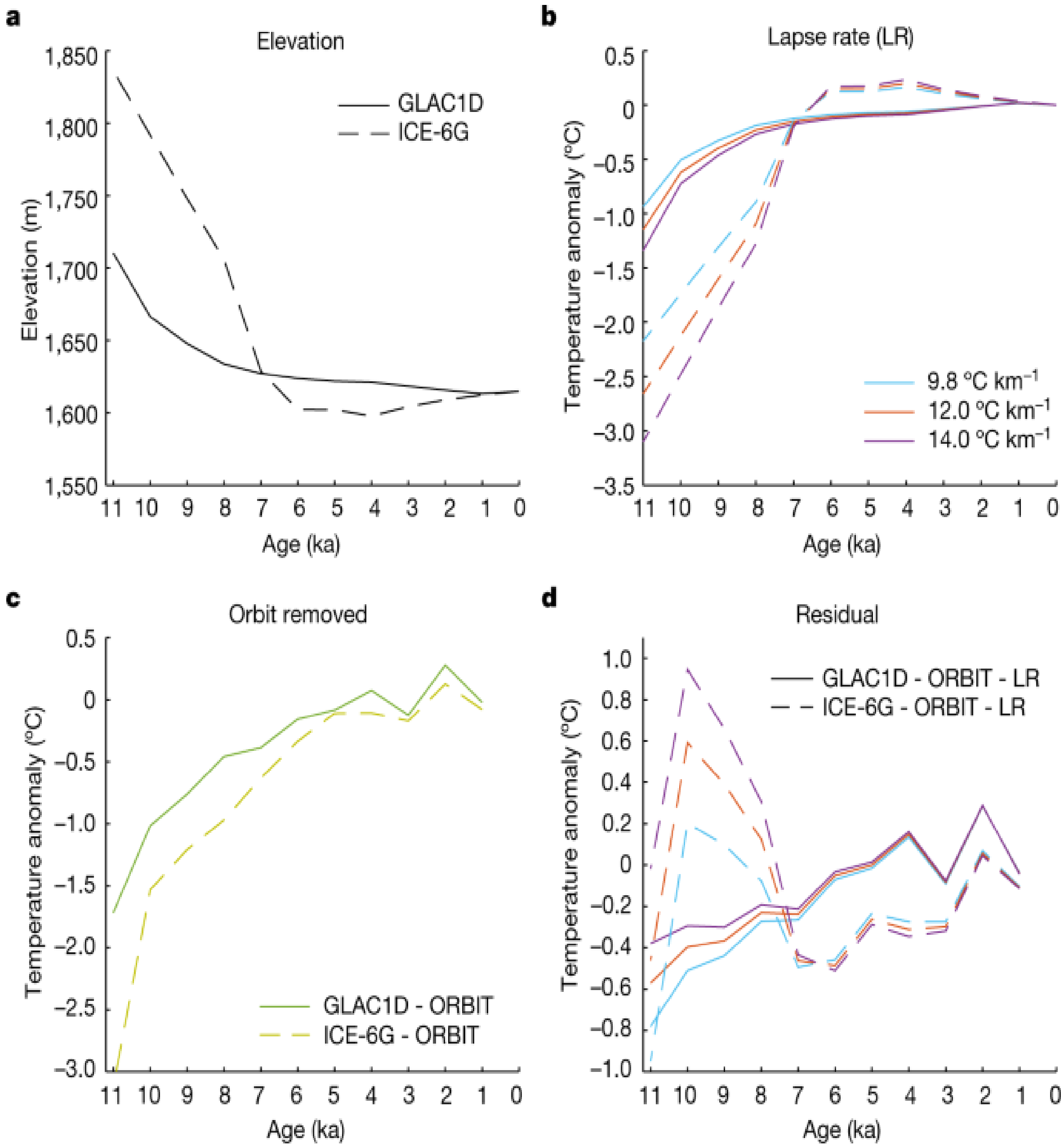
HadCM3 simulations

To investigate the role of more-complex geography and mechanisms, including topographical changes not accounted for in the MEBM, we simulated Holocene climate with a fully coupled general circulation model, HadCM3 (ref. [35]) (Methods—HadCM3 model simulations). Simulations forced solely by changes in orbital parameters produce summer maximum temperatures (for approximately the December solstice) at 80° S similar to our reconstructed values and to the MEBM: increasing over the Holocene, peaking at 4 to 3 ka and decreasing into the modern era (ORBIT, Fig. 2a). This pattern reflects a strong role of maximum summer insolation in determining observed summer temperatures. The similarity of the early- to mid-Holocene (11–6 ka) summer temperature increase in the orbitally forced HadCM3 simulations and our reconstruction suggests little influence of changing ice-sheet elevation and extent. A similar comparison for winter yields an approximately 1.25 °C decrease of model ORBIT temperature

(Fig. 2c) compared to a possible small increase in temperature in the reconstruction (Fig. 2d; >90% chance of >0.1 °C), suggesting some warming resulting from a lowering ice sheet.

Next, as boundary conditions in the HadCM3 simulations, we prescribed variable greenhouse gas (GHG) concentrations and two different ice-sheet histories, GLAC1D and ICE-6G, which entail net surface lowerings of about 83 m and about 208 m, respectively, from 11 to 7 ka at the WDC site (Fig. 4a). These elevation scenarios substantially affect simulated temperatures (Fig. 2a,c). Much of the elevation-induced warming in these models, which occurs primarily in the early Holocene, can be attributed directly to the surface lapse-rate effect (Fig. 4b). However, comparison to the orbital-only runs (Fig. 4c) reveals a remaining temperature anomaly (Fig. 4d), attributable to GHGs, ice-sheet extent and nonlinear responses to simultaneously imposed forcings. Sea ice has only a small impact on the temperature at 80° S in summer (Methods—Sea ice; Extended Data Fig. 6).

Fig. 4: Possible ice elevation histories and the corresponding modelled temperatures.



a, Elevation histories used in HadCM3. GLAC1D[43] is 96 m higher at 11 ka compared to present and ICE-6G[44] is 222 m higher. **b**, Temperature anomalies from elevation change (GLAC1D, solid lines; ICE-6G, dashed lines) using an atmospheric lapse rate of 9.8 °C km^{-1} and spatial lapse rates for interior West Antarctica of 12 °C km^{-1} (ref. [45]) and 14 °C km^{-1} (ref. [46]). **c**, HadCM3 residual-temperature anomalies for December (summer) calculated by subtracting the ORBIT run from the GLAC1D and ICE-6G runs in Fig. 2a, highlighting the portion attributable to changing elevation

rather than insolation. **d**, Residual-temperature change in **b** subtracted from the results in **c**, showing the component driven from processes besides the direct lapse-rate (LR) effect and orbital forcing.

Source data

Inconsistencies exist between the different ice-sheet scenarios (Fig. 4d) and the summer versus winter seasons but differences are minor enough to permit a bounded estimate of the true Holocene elevation decrease. This calculation is made by comparing the excess of the reconstructed temperature increase over the orbital-only simulation to the same excess for the ice-sheet model simulations and scaling to the elevation changes used in the latter (Methods—Estimating elevation changes). We find central estimates for elevation decrease of 23 m and 53 m from comparison to the GLAC1D and ICE-6G scenarios, respectively, over the period 10 to 3.5 ka (Table 1). Accounting for uncertainties in the seasonal temperature reconstructions (Fig. 2) allows for elevation changes ranging from 33 m increase to 131 m decrease (2σ) from 10 to 3.5 ka or 54 m increase to 162 m decrease (2σ) if the time interval is narrowed to 10 to 6.5 ka (Table 1). Our results, thus, are consistent with geological observations of ice high-stands on mountain nunataks, which indicate less than 100 m of Holocene surface lowering[4,5,6].

Table 1 WAIS elevation decrease

Winter temperatures on the Antarctic mainland must respond to insolation forcings indirectly, via heat transport from lower latitudes. Orbital forcing models predict winter cooling across the Holocene, mostly from 11 to 6 ka (Figs. 2c and 3g). Both models and reconstructed winter temperatures lack a late Holocene maximum. But in the earlier Holocene, the winter reconstruction does not display the cooling trend expected from models and is dominated by prominent millennial variations. The mismatch with insolation at lower latitudes and absence of local forcings suggests variations in the efficacy of meridional atmospheric-heat transport.

Discussion

Diverse and numerous proxies are used to reconstruct globally averaged surface temperatures for evaluating climate models and distinguishing natural from anthropogenic climate variability[33,34,36,37,38]. How these proxies depend on seasonal factors has been assessed in only a few cases[39]. Our West Antarctic study provides a cautionary example, as the mean annual temperature history reflects different controlling factors of summer and winter temperatures whose importance varies with time. In such a situation, important seasonal dynamics may be missed, or proxies misinterpreted, when only mean climate is considered. In addition, incorporating more information from the southern polar regions should help global temperature assessments avoid biases associated with weighting of temperature reconstructions toward northern sites, which have produced differing interpretations of the relationship between global climate and forcings in the Holocene, even including opposing trends[34,40,41].

Previous analyses with simplified atmospheric models[3] identified the duration of Southern Hemisphere summer as a key driving variable of Antarctic climate at orbital timescales. Some palaeoclimate findings validate this claim; for example, the onset of deglacial warming in West Antarctica corresponds with increasing integrated summer insolation[2]. Our results—spanning about half a precession cycle—reveal a dominant role for annual maximum insolation in determining West Antarctic summer climate during the Holocene, without precluding a greater role for duration or integrated summer insolation in other periods, such as glacial terminations.

Methods

We measured WAIS Divide core (WDC) water isotopes using continuous-flow analysis (see next section on Water isotopes) and then corrected for cumulative diffusion using spectral techniques to determine diffusion lengths and restore prediffused amplitudes within 140-yr sliding windows (section below on Diffusion corrections; Extended Data Fig. 1c). Summer maxima and winter minima (Fig. 1a) identified in these corrected data were then used to calculate summer and winter amplitudes for each year. We converted the isotope amplitudes to temperature amplitudes using a model-determined scaling factor (section below on Seasonal temperatures) and added them to previously reconstructed mean annual temperatures[30] to

recover summer and winter values. Substantial seasonal noise processes required multicentennial to millennial averaging to reduce uncertainty (section below on Uncertainties in reconstructing temperatures). To elucidate physical controls on subannual temperatures, we used a simple energy balance model and HadCM3, a general circulation model, to calculate expected changes in seasonal and monthly surface temperatures through time under varying boundary conditions (sections below on Seasonal moist energy balance model and HadCM3 model simulations). Finally, using both observations and modelling, we estimated the change in WAIS surface elevation through the Holocene (section below on Estimating elevation changes).

Water isotopes

WDC water isotopes (Extended Data Fig. 1a) were analysed on a continuous-flow analysis system[21] using a Picarro cavity ring-down spectroscopy instrument, model L2130-*i*. Using permutation entropy[47], we identified data anomalies arising from laboratory analysis, which were corrected, including by resampling ice through 1,035.4–1,368.2 m depths (4,517–6,451 yr)[48]. All other Holocene data are previously published[18,31] and available online[19,20]. Data are reported in 5-mm increments in delta-notation (‰, or per mil) relative to Vienna standard mean ocean water ($\delta^{18}O = \delta D = 0$‰), normalized to standard light Antarctic precipitation ($\delta^{18}O = -55.5$‰, $\delta D = -428.0$‰). WDC is annually dated, with accuracy better than 0.5% of the age between 0 and 12 ka (ref. [22]). For the Holocene, the temporal spacing of consecutive 5 mm samples is <0.1 yr and the average <0.05 yr, ranging from about 2.6 weeks at 10 ka to 0.5 week at 1 to 0 ka (ref. [18]).

Diffusion corrections

Diffusion in the firn and deeper ice attenuates high-frequency water-isotope information in ice cores[26,27,49,50,51,52]. Diffusion length quantifies the statistical vertical displacement of water molecules from their original position[27,49]. We used diffusion-correction code developed by S. Johnsen, University of Copenhagen[23,24,27,28], which uses maximum entropy methods

to invert an observed power-density spectrum. As an input to these inversions, we determined diffusion lengths (Extended Data Fig. 1c) for 140-yr windows using previous methods[18,26]. The power-density spectrum observed in the ice-core record $P\left(f\right)$,after diffusion, is $P\left(f\right)={P}_{o}\left(f\right){\rm{\exp }}\left[-{\left(2\pi f{\sigma }_{z}\right)}^{2}\right]$, in which ${P}_{o}\left(f\right)$ represents the power spectrum of the undiffused signal (‰^{2} m^{-1}), f is the frequency $\frac{1}{\lambda }$ (1/m), λ the signal wavelength (m), *z* the depth (m) and ${\sigma }_{z}$ the diffusion length (m). The original, prediffusion power-density spectrum (diffusion-corrected) is calculated as ${P}_{o}\left(f\right)=P\left(f\right){\rm{\exp }}\left(4{\pi }^{2}{f}^{2}{\sigma }_{a}^{2}\right)$, for diffusion length ${\sigma }_{a}$ (yr) and f now with units of 1/yr. The ${\sigma }_{a}=\frac{{\sigma }_{z}}{{\lambda }_{{\rm{avg}}}}$, in which ${\lambda }_{{\rm{avg}}}$ is the mean annual layer thickness (m yr^{-1}) at a given depth. The diffusion-corrected spectrum takes the form of a series of complex numbers ${X}_{{\rm{R}}}+i{X}_{{\rm{I}}}$ versus f. From this, the amplitude spectrum A is obtained by $A\left(f\right)=\sqrt{{X}_{R}^{2}+{X}_{{\rm{I}}}^{2}}$ and the phase spectrum φ is obtained by $\varphi \left(f\right)={{\rm{\tan }}}^{-1}\left(\frac{{X}_{{\rm{I}}}}{{X}_{{\rm{R}}}}\right)$. The real components of the amplitude and phase spectrums give the diffusion-corrected water-isotope signal ${\delta }_{o}\left(t\right)$ as:

$${\delta }_{o}\left(t\right)=\mathop{\sum }\limits_{i=1}^{N}{A}_{i}{\rm{\cos }}\left(2\pi {f}_{i}t+{\varphi }_{i}\right)$$

Uncertainties on ${\delta }_{o}\left(t\right)$ are determined using the uncertainty range for diffusion lengths[26] calculated in each 140-yr window. Before spectral analysis, the isotope data are linearly interpolated at a uniform time interval of 0.05 yr. Our determination of diffusive attenuation and correction arises from the observed frequency spectra themselves and therefore is entirely independent of firn diffusion and densification models.

Seasonal water-isotope amplitudes

To select extrema (summers and winters) in the diffusion-corrected δD signal (Fig. 1a and Extended Data Fig. 1b), we used the 'findpeaks' MATLAB function. Figure 1c,e show the resulting time series for summer and winter, averaged with a 50-yr boxcar filter for clarity of trends. For every year defined in the WDC age-scale, we calculated the averaged diffusion-corrected δD. The difference between the two extrema and the mean define the summer and winter isotope amplitudes.

Seasonal temperatures

A linear scaling converted seasonal isotopic amplitudes to seasonal temperature amplitudes, using a sensitivity of isotopes to surface temperatures determined by the simple water-isotope model (SWIM)[29]. Finally, to find summer and winter temperatures we added the individual seasonal temperature amplitudes to the year's mean temperature obtained previously[30] by calibrating the water-isotope record against borehole temperatures and $\delta^{15}N$ constraints on firn thickness.

SWIM is based on earlier numerical Rayleigh-type distillation models[53,54], which simulate the transport and distillation of moisture down climatological temperature gradients. As moist air is transported towards the poles and cools, the saturated vapour pressure decreases nonlinearly and moisture above saturation is removed by precipitation. The model keeps track of the isotopic fractionations at each step along this distillation process. In most previous simple models, there is an inconsistency in the calculation of the supersaturation that determines the point of condensation and that drives kinetic isotope fractionation. Modifications to these earlier models, used in SWIM, ensure consistency in the calculation, which results in a smoother relationship between temperature and the δ-values of precipitation and better agreement with observed spatial patterns of δD and $\delta^{18}O$. Given input of both δD and $\delta^{18}O$ data, SWIM calculates distributions of source temperatures, the temperature gradients of pseudo-adiabatic pathways and condensation temperature. We used SWIM to derive sensitivities for surface isotope-temperature scalings using diffusion-corrected WDC data to obtain a surface scaling of 6.96‰ δD $°C^{-1}$. Using raw data, the surface scaling is 7.07‰ δD $°C^{-1}$. In comparison to other isotope-temperature scalings, ref. [55]

obtain about 6.56‰ δD °C^{-1} and ref. [30] about 7.10‰ δD °C^{-1} (both converted from $\delta^{18}O$ to δD using a factor of 8).

Uncertainties in reconstructing temperatures

We included uncertainties associated with the following factors: measurement analysis, diffusion correction, seasonality of accumulation, precipitation intermittency, modelled isotope-temperature scaling and mean-temperature history. The 'analysis uncertainty' is 0.55‰ for δD (1σ) (ref. [21]). The 'diffusion-correction uncertainty' is described in ref. [26]. The uncertainty of the mean-temperature reconstruction, calculated previously[30], accounts for most uncertainty in the early Holocene but a small fraction in the late Holocene. Sections 'Seasonal preservation bias uncertainty' to 'Isotope-temperature scaling and associated uncertainty' below explain the other uncertainty terms. Uncertainties for some factors (analysis and diffusion correction) can be treated as independent random variables so that, on time-averaging, their magnitudes decrease as the inverse of the square root of the number of values. Uncertainties for other factors (intermittency, isotope-temperature scaling, mean temperature and seasonality) might be systematically biased and therefore their magnitudes are taken to be invariant with respect to the interval of averaging. On the basis of the 2σ uncertainties for summer and winter temperature (Fig. 2a,c), we assessed the significance of dominant trends using Monte Carlo analysis (Fig. 2b,d; section on Trend analysis below).

Seasonal preservation bias uncertainty

Unequal seasonal distribution of snowfall could result in different magnitudes of diffusion for winter and summer amplitudes[49]. The seasonal temperature cycle also affects the magnitude of diffusion for all seasons. We used the Community Firn Model (CFM)[56,57], a firn-evolution model with coupled firn temperature, firn densification and water-isotope modules, to test how seasonally weighted accumulation affects the diffusion of specified, hypothetical isotope records progressing from surface snow (δD_{snow}), to consolidated snowpack in the firn (δD_{firn}), to solid ice beneath the pore close-off depth (δD_{ice}). We applied the back-diffusion calculation (section on

Diffusion corrections) to δD_{ice} to estimate the original δD_{snow}. We then assessed how reconstructions of δD_{snow} could be misinterpreted as a result of different seasonal-accumulation weightings (Extended Data Fig. 2a,b).

We performed five CFM runs using monthly time steps for accumulation, temperature and isotopes (Extended Data Table 1). The seasonal cycle for δD_{snow} is based on the mean amplitude in Fig. 1b (15.43‰). Five WAIS accumulation scenarios were tested on the basis of monthly accumulation from the regional climate model MAR3.6 (Modèle Atmosphèrique Règional; ERA-Interim forced)[58], which spans the period January 1979to December 2017. The mean accumulation over the entire 39-yr period is 0.24745 m ice equivalent yr^{-1}, with about 1.6× as much snow in winter (April to September) as in summer (October to March). The five scenarios are as follows: (1) ‘constant’: identical accumulation for all months (0.0206 m ice equivalent $month^{-1}$; one-twelfth of the annual mean); (2) ‘cycle’: monthly accumulation equal to MAR monthly means; (3) ‘noise’: using the ‘cycle’ time series, we add noise to each time step in the ‘cycle’ series in the form of a normal random variable of zero mean and the standard deviation for the month from MAR; (4) ‘random’: for each month, the accumulation is a normal random variable with mean and standard deviation equal to MAR monthly values; and (5) ‘loop’: the entire 39-yr MAR accumulation time series is repeated over and over again. For the temperature boundary condition, we used the mean temperature at a height of 2 m for 1979–2017 for each month predicted by MAR to create an annual temperature cycle. We repeated this 12-month time series for the duration of the model runs. This method ensures that model runs, which are designed to test accumulation seasonality, are not affected by interannual temperature variability, while also providing an estimate of the annual temperature cycle, which affects the rate of isotope diffusion in the upper firn.

Extended Data Fig. 2a,b shows the results for the ‘constant’ and ‘cycle’ cases. The diffusion-correction technique accurately reconstructs δD_{snow} for summer and winter in the ‘constant’ snowfall scenario but underestimates summer values in the MAR ‘cycle’ scenario by about 2.6‰, which is 11% of the full range of the observed WDC summer water-isotope values. Winter values are overestimated by only about 0.6‰, about 3% of the full winter range, since winter has 1.6× as much snow as summer. The ‘noise’,

‘random’ and ‘loop’ runs produce results within 0.3‰. These CFM experiments demonstrate that centennial trends in the summer and winter water isotopes of the order of a few per mil (‰) could arise from large changes in seasonal-accumulation weighting, whereas multimillennial trends ≫2.6‰ are unlikely to be caused by seasonal accumulation and can therefore be interpreted as climate signals of a different origin. For 1,000-yr averaging (as in Fig. 2), HadCM3 indicates seasonal-accumulation weighting (winter:summer) of 1.3 to 1.7 throughout the Holocene (Extended Data Fig. 2f), which yields a 1σ uncertainty of 0.27‰ based on the CFM testing criteria.

To determine observationally if seasonal snowfall changed across the Holocene, we used measured black carbon (BC) concentrations, the only age-scale-independent impurity. BC data are available from 0–2.5 ka and 6–11 ka (ref. [22]). Seasonal-fire regimes in South America dominate BC concentrations at WDC, causing BC maxima and minima in autumn and spring, respectively[59] (Extended Data Fig. 2c). We split each year into two parts, characterized by rising or falling BC: BC_1 and BC_2 the depth intervals of rising and falling BC (Extended Data Fig. 2d). The duration of BC_2 is longer than BC_1 owing to source characteristics[59], thus $BC_1/BC_2 < 1$ (Extended Data Fig. 2e). The BC_1/BC_2 ratio can change with time because of variability at the source, changes in atmospheric transport or seasonality of snow deposition. We observe little change in BC_1/BC_2 resembling the multimillennial trends seen in WDC summers and winters (Extended Data Fig. 2e). Unless there are competing and exactly compensating effects in seasonality (the source change exactly cancels the depositional and transport change or other unlikely scenarios), the BC data provide evidence that changes in WDC seasonal snowfall were not large enough to affect our multimillennial climate interpretations.

Intermittency of precipitation uncertainty

The episodic nature of snowfall creates an incomplete record of local climate variations[60], preventing interpretation of trends over short time intervals. We want to interpret isotopic variations averaged over a sufficiently long timescale so that, to within a specified tolerance, trends are

not likely to be random noise arising from the spread of distributions preserved in the ice. Using distributions of reconstructed annual amplitudes (Fig. 1b and Extended Data Fig. 2g) for 1,000-yr windows throughout the Holocene, we conducted Monte Carlo resampling simulations to determine that 250-yr averaging-lengths are needed to achieve a standard error of 1‰, corresponding to a mean amplitude-to-noise ratio of 15. For the time period with greatest variability, centred on 4 ka, the standard error for a 1,000-yr average (as used in Fig. 2) is 0.52‰ (Extended Data Fig. 2h). Because this is an amplitude uncertainty (rather than uncertainty associated with a season), we specify the 1σ uncertainties for summer and winter as half of 0.52‰.

Isotope-temperature scaling and associated uncertainty

The conversion of isotopic values (1,000-yr averages) to temperature yields three curves for summer and three for winter: T_{nominal}, $T_{+1\sigma}$ and $T_{-1\sigma}$. Each curve is normalized to the value at 1 ka (as done in Fig. 2). The difference in the $T_{+1\sigma}$ and $T_{-1\sigma}$ curves gives the 1σ uncertainty range $+\sigma_{\mathrm{Tscale}}$ to $-\sigma_{\mathrm{Tscale}}$, which are then added in quadrature to the ref. [30] mean-temperature uncertainties, yielding the final uncertainty estimates shown in Fig. 2.

Relationship between the annual mean and individual seasons

Using 1,000-yr and 300-yr averages of summer, winter and mean temperature (Extended Data Fig. 3c–f), we determined R^2 values for summer and winter versus the mean. We then subtracted the 1,000-yr averages from the 300-yr averages to obtain residuals and then determined R^2 values again for summer and winter versus the mean (Extended Data Table 2). From 11 to 0 ka, the high summer correlations for the 1,000-yr comparison indicate a strong association of the annual mean temperature with the summer temperature at orbital timescales. At suborbital scales (300–1,000-yr residuals), neither the summer nor the winter alone explain much of the mean annual variability and the annual mean is a random composite of the two seasons. If only 11–7 ka is considered, winter variability explains more of the mean at submillennial scales.

Trend analysis

To assess the significance of dominant trends in our reconstructed seasonal temperatures, we conducted a Monte Carlo analysis founded on the assumption that all possibilities for the unknown time-dependence of errors are equally likely. The essentialmotivation for this approach is that we have determined the magnitudes of uncertainties as a function of age but that we have no information about whether the errors in our reconstruction persist at similar values for long periods of time (exhibit a bias) or whether they fluctuate at high-frequency.

We randomly generated a large number of alternative seasonal temperature histories governed by the uncertainties on 1,000-yr averages (Extended Data Fig. 3a,b), calculated the temperature trends for each alternative history over desired time intervals (such as 11–4 ka and 4–0 ka) and compiled the results into frequency distributions from which probabilities can be calculated (Fig. 2b,d). Specifically, each alternative history deviates from the summer and winter temperature reconstruction by an amount that smoothly varies over time between random nodes whose values are a Gaussian random variable of zero mean and standard deviation for 1,000-yr averages at the age of the node. The number of nodes and the age of each node are random variables, uniformly distributed between 1 and 11 nodes and 0–11 ka, respectively. A small number of nodes produce an alternative temperature history for which the bias is serially correlated for millennia, whereas a large number of nodes produce a history for which the bias is uncorrelated from millennium to millennium.

Seasonal moist energy balance model

We used a simple global, zonal mean MEBM to calculate surface temperatures (Extended Data Fig. 4), accounting for TOA insolation, temperature-dependent longwave emission to space, temperature-dependent albedo to simulate brightening by snow and ice and horizontal atmospheric-heat transport treated as diffusion of near-surface moist static energy[61,62,63]. The model has a 2° spatial resolution, a single surface and single atmospheric layer and a surface-heat capacity based on the relative fraction of land and ocean surface in the zonal mean. Heat exchange between the surface and atmosphere layers arises from differences in blackbody radiation from each layer and sensible and latent heat exchanges proportional to

temperature and specific humidity contrasts (assuming a constant relative humidity of 80%), following bulk aerodynamic formulae. We calculated the annual TOA insolation-cycle at every latitude in 500-yr time slices from 0 to 11 ka. For each time slice the model is run at 2 hour time resolution for 30 model-years to reach equilibrium.

Extended Data Fig. 4 compares the temporal evolution of summer maximum and winter minimum heat divergence by the atmosphere ($\nabla \cdot \mathbf{F}$) at the WDC site to the summer maximum and winter minimum site temperature and insolation. Although Holocene changes in $\nabla \cdot \mathbf{F}$ at the WDC site correlate with insolation forcing, the magnitude of changes in maximum direct insolation are much larger than those in atmospheric-heat divergence. Further, the Holocene changes in summertime heat divergence are of the wrong sign to cause net heating at the WDC site (positive divergence is an export of heat by the atmosphere from the site). Heat transport in the Antarctic is convergent in the annual mean but divergent in mid-summer, as intense incoming insolation exceeds longwave emissions from the cold surface.

HadCM3 model simulations

Model setup

We used the fully coupled ocean-atmosphere model HadCM3 (refs. [64,65]), v.HadCM3BM2.1, which well simulates tropical Pacific climate and its response to glacial forcing[66]. Our simulations are snapshots at 1 kyr intervals over the last 11 ka (ref. [35]), with time-specific boundary conditions of orbital forcing[67], GHG concentration[68,69], ice-sheet topography and sea level[43,44,70,71,72,73,74]. We used three simulations: (1) only orbital forcing changes (ORBIT), with all other boundary conditions set to the pre-industrial; (2) orbital/GHG forcing with GLAC1D ice-sheet elevation history; and (3) orbital/GHG forcing with ICE-6G ice-sheet elevation history. Elevation histories are shown in Fig. 4a. Snapshot simulations were run for at least 500 yr with analysis made on the final 100 yr. Further snapshot simulations for 10 ka allowed us to decompose the role of different forcings, described in the following sections. The large difference in

forcings between 10 ka and the pre-industrial late Holocene epoch make this comparison most instructive.

Summer climate

We examined the zonal mean at 80° S in simulations for hypothetical 10 ka worlds, by changing the boundary conditions to compare to the pre-industrial/late Holocene. These simulations are '10 ka ORBIT-only'; two runs with only ice sheets at 10 ka and pre-industrial settings otherwise, called '10 ka GLAC1D-only' and '10 ka ICE-6G-only'; and two runs with all 10 ka forcings, called '10 ka GLAC1D-all' and '10 ka ICE-6G-all'. In '10 ka ORBIT-only', reduced TOA shortwave radiation causes a large reduction in shortwave radiation at the surface (SW_d) and consequent cooling. Downward longwave radiation (LW_d) also decreases, probably because of atmospheric cooling. Sensible heat flux (SH_d) to the surface is increased, indicating that the atmosphere and surface do not equally cool; one cause of this is the increased meridional heat convergence ($-\nabla\cdot\mathbf{F}$).

Changed ice sheets cause summertime cooling in both '10 ka GLAC1D-only' and '10 ka ICE-6G-only', primarily via reduced LW_d. Increased SW_d, as a result of reduction of depth of the atmospheric column above the ice-sheet surface, counteracts the reduced LW_d to some extent. (Reducing the atmospheric column reduces SW absorption and tends to cool the atmosphere, reducing LW_d). Both ice-sheet scenarios also cause an increase in $-\nabla\cdot\mathbf{F}$, partly counteracting summertime cooling.

Using all 10 ka forcings causes cooling through both LW_d and SW_d. The decrease in SW_d is similar in '10 ka GLAC1D-all' and '10 ka ICE-6G-all' and slightly smaller than in '10 ka ORBIT-only', probably because the thinner atmospheric column reduces absorption. The decrease in LW_d in '10 ka GLAC1D-all' and '10 ka ICE-6G-all' is larger than in '10 ka ORBIT-only', '10 ka GLAC1D-only' and '10 ka ICE-6G-only'. This indicates the importance of feedbacks in the atmosphere. Heat convergence $-\nabla\cdot\mathbf{F}$ increases in both simulations indicating remote feedbacks, in addition to local feedbacks related to the amount of water vapour in the atmosphere.

The preceding description of changes in the zonal mean in the 10 ka simulations compared to pre-industrial period holds for the entire Holocene epoch. Orbital forcing alone reduces SW_d and LW_d by roughly the same magnitude. With full forcing (including ice sheets), the reduction in LW_d is roughly three times the reduction in SW_d. Considering an energy budget over the WDC site (79.467° S, 112.085° W), mechanisms are the same as for the zonal mean. Magnitudes of forcings change but reduced SW_d still cools the surface, amplified by an LW_d feedback dependent on ice-sheet size.

Winter climate

During winter, SW_d is no longer a factor as the sun is below the horizon, yet there is still surface warming caused by an increase in LW_d. An increase in $-\nabla\cdot\mathbf{F}$ in ‘10 ka ORBIT-only’ warms the atmospheric temperature, increasing LW_d and SH_d. With an ice sheet imposed, the surface temperature cools. In both ‘10 ka GLAC1D-only’ and ‘10 ka ICE-6G-only’ there is a reduction in $-\nabla\cdot\mathbf{F}$, reducing LW_d and SH_d. When all 10 ka forcings are introduced, the change in temperature is smaller than for ice sheet-only runs. In ‘10 ka GLAC1D-all’ we found no change in $-\nabla\cdot\mathbf{F}$, LW_d or surface temperature. This suggests that the increase in $-\nabla\cdot\mathbf{F}$ from orbital forcing is almost perfectly balanced by the change in $-\nabla\cdot\mathbf{F}$ from the ice-sheet configuration.

The processes controlling heat transport over Antarctica are complicated and HadCM3 may not be able to simulate them perfectly. Our simulations indicated that remote processes during winter alter the heat transport, affecting atmospheric and surface temperatures. Raising the topography of Antarctica tends to reduce such heat transport (Extended Data Fig. 5), producing an additional cooling on top of a pure lapse-rate effect[75]. This cannot, however, explain the prominent millennial-scale changes at about 9.2 and about 7.9 ka (Figs. 1e and 2c). The intricacies of interpreting the early Holocene winter variability in West Antarctica necessitates further study.

Sea ice

Sea ice changes may alter local energy fluxes from the ocean to the atmosphere. In HadCM3, sea ice extent changes across the Holocene. We used two analyses (Extended Data Fig. 6) to show that sea ice is not a primary control on the surface temperature at WDC (80° S): (1) correlation analysis of sea ice changes and surface temperature and (2) atmosphere-only model simulations in which we specified individual changes in the model boundary conditions (including sea ice).

We computed the dominant spatial patterns of sea ice variability using empirical orthogonal functions across all of the HadCM3 simulations (ALL) and individually for three subset simulations (ORBIT, GLAC1D and ICE-6G), from 0 to 11 ka. We projected the model-simulated sea ice for each individual time-slice simulation onto these patterns to compute the amplitude of sea ice variability in each simulation. The amplitude was compared to temperature at 80° S to understand how large-scale changes in the sea ice affect temperature for the months of December (summer) and July (winter) (Extended Data Fig. 6).

In winter, we found negligible correlations between sea ice change and temperature in all sets of simulations (ALL: 0.02; ORBIT: 0.04; GLAC1D: 0.16; ICE-6G: 0.06). This suggests that winter sea ice is not an important factor in determining the temperature at 80° S. In summer, only the ORBIT simulation has meaningful correlations between temperature and sea ice variability (ALL: 0.59; ORBIT: 0.84; GLAC1D: −0.60; ICE-6G: 0.36). The sign of the correlation changes between simulations despite the sea ice change pattern being the same in all simulations. From this we concluded that sea ice is not a dominant control on temperature at 80° S in summer. The correlation in the ORBIT simulations suggests that there may be some relationship between sea ice and temperature; we investigated this with atmosphere-only simulations.

In atmosphere-only simulations at 0 ka and 10 ka, we specified the top of the atmosphere insolation, sea surface temperature (SST) and sea ice from the ORBIT simulations. Over land areas and sea ice regions, the model calculates the surface temperature using the land-surface scheme in the model. The atmosphere model is identical to the model used within the coupled model. We ran a series of experiments varying the orbital configuration, SST or sea ice (summarized in Extended Data Table 3).

The zonal mean of the change in the sea ice that we prescribed can be seen in Extended Data Fig. 6e and the change in the SST can be inferred from Extended Data Fig. 6f–h. Extended Data Fig. 6f shows that the atmosphere-only model replicates the change in temperature of the coupled model. Extended Data Fig. 6g shows that the effect of the 10 ka orbital configuration ('Atmos_10k_insol') is to cool Antarctica considerably by about 0.5 °C. North of 65° S there is no change in the surface temperature, primarily a response to the imposed SST and sea ice, which are the same in the 'Control' and 'Atmos_10k_insol'. Imposing the SST and sea ice from 10 ka ('Atmos_10k_ice_SST'), we find very little change in the surface temperature over Antarctica but there are some large changes in the surface temperature north of 70° S. Extended Data Fig. 6h shows the result of imposing 10 ka SST or 10 ka sea ice. The 10 ka SST ('Atmos_10k_SST') tends to warm Antarctica, consistent with the large increases in SST north of 65° S. Changing sea ice ('Atmos_10k_ice') tends to cool Antarctica. Both effects are small, approximately 0.1 °C and of opposite sign. This explains the small net change in the surface-temperature change over Antarctica when SST and sea ice are changed simultaneously, as shown by 'Atmos_10k_SST_ice'. It should be noted that in the coupled system a change in sea ice cannot be decoupled from a change in the SST, so not only is the effect of sea ice on the climate small, it is also probably associated with a compensating change in SST. From these simulations we concluded that sea ice and SST changes are not a dominant driver of the change in the surface temperature over Antarctica.

Extended Data Fig. 6e shows that the change in sea ice at 10 ka in ORBIT is much larger than the change in either GLAC1D or ICE-6G. The ORBIT simulations do not account for all of the changes in the boundary conditions at 10 ka and are therefore less realistic than either ICE-6G or GLAC1D. Because ICE-6G and GLAC1D both show much smaller changes in the sea ice and SST than ORBIT, we expect that in reality there is also a much smaller change in the sea ice and SST than in ORBIT. We thus concluded that sea ice has a small impact on the temperature at 80° S in summer.

We also performed a similar analysis of the winter season (not shown). We found that the atmosphere-only model does not compare well with the coupled model, simulating very little change in the surface temperature. Doing a term-by-term decomposition of the atmosphere model is not,

therefore, particularly useful as it tells us more about the model rather than the physical climate. The failure of the atmosphere-only model to capture the changes at 10 ka suggests that the importance of the SST and sea ice is in their day-to-day coupling with the atmosphere and not in any long-term mean change in this season.

Estimating elevation changes

Temperatures simulated by HadCM3 for ORBIT provide a control scenario against which observations can be compared to identify the signal of elevation change. For a chosen time interval, the net reconstructed warming ΔT_R exceeds that of ORBIT by an amount $\Delta T_R - \Delta T_O$. This can be compared to the effective lapse rate $(\Delta T_M - \Delta T_O)/\Delta Z_M$ defined by a HadCM3 simulation including topographic change (GLAC1D or ICE-6G models) and all forcings, for model warming ΔT_M and model elevation decrease ΔZ_M. Specifically, the estimated elevation decrease is $\Delta Z_R = \Delta Z_M((\Delta T_R - \Delta T_O)/(\Delta T_M - \Delta T_O))$. Summer and winter reconstructions offer two separate assessments, for which we calculate the algebraic average.

Accounting for uncertainties in ΔT_R requires recognizing that uncertainties of summer and winter reconstructions are not independent, while also recognizing that they emerge from two independent sources: uncertainty in the mean annual temperature history (calculated in ref. [30]) and uncertainty in the seasonal amplitude (calculated in the present study). In general, the uncertainty of seasonal temperature at a specified time is the quadrature sum of annual and amplitude uncertainties, which gives 1σ and 2σ uncertainties for a given season and time. However, if the true value of annual temperature is shifted by an amount $\alpha\sigma$ from the nominal reconstruction, this must be true for both summer and winter. And if the true value of amplitude is shifted by an amount $\beta\sigma$ from the nominal reconstruction, the temperature shift must be $+\beta\sigma$ in one season but $-\beta\sigma$ in the other.

To define bounding cases on elevation change in a specified time interval, we calculated the maximal (or minimal) temperature change ΔT_R for a season by differencing the upper (or lower) limit at one end of the interval with the lower (or upper) limit at the other end and also calculating the

corresponding ΔT_R for the opposite season required by the correlated errors. The elevation decreases ΔZ_R were then calculated by comparison to HadCM3 simulations, as specified previously and the summer and winter values averaged. This process was completed four times for each time interval and HadCM3 model, corresponding to four different initial ΔT_R (maximum and minimum ΔT_R for summer and maximum and minimum ΔT_R for winter) and the most extreme case taken as the result (this proved to be the one starting with maximum summer ΔT_R). Table 1 lists results for two time intervals and the two HadCM3 simulations with variable topography.

Data availability

The WDC water-isotope datasets analysed during the current study are available in the US Antarctic Program Data Center (USAP-DC) repository, https://doi.org/10.15784/601274 and https://doi.org/10.15784/601326. The impurity datasets analysed during the current study are available in USAP-DC repository, https://doi.org/10.15784/601008. The data generated in this study are available in the USAP-DC repository, https://doi.org/10.15784/601603, including raw and diffusion-corrected water isotopes, seasonal water isotopes (maximum summer and minimum winter values) and seasonal temperature reconstructions. Source data are provided with this paper.

Code availability

The MATLAB code used for the diffusion correction of water-isotope data and the subsequent selection of seasonal extrema (summer and winter) are available online at Zenodo, https://doi.org/10.5281/zenodo.7042035.

References

1. Berger, A. Milankovitch theory and climate. *Rev. Geophys.* **26**, 624–657 (1988).

2. WAIS Divide Project Members. Onset of deglacial warming in West Antarctica driven by local orbital forcing. *Nature* **500**, 440–444 (2013).

3. Huybers, P. & Denton, G. Antarctic temperature at orbital timescales controlled by local summer duration. *Nat. Geosci.* **1**, 787–792 (2008).

4. Ackert, R. P. et al. Measurements of past ice sheet elevations in interior West Antarctica. *Science* **286**, 276–280 (1999).

5. Steig, E. J. et al. in *The West Antarctic Ice Sheet: Behavior and Environment* Vol. 77 (eds Alley, R. B. & Bindschadle, R. A.) 75–90 (American Geophysical Union, 2001).

6. Ackert, R. P., Mukhopadhyay, S., Parizek, B. R. & Borns, H. W. Ice elevation near the West Antarctic Ice Sheet divide during the last glaciation. *Geophys. Res. Lett.* https://doi.org/10.1029/2007GL031412 (2007).

7. Spector, P., Stone, J. & Goehring, B. Thickness of the divide and flank of the West Antarctic Ice Sheet through the last deglaciation. *Cryosphere* **13**, 3061–3075 (2019).

8. Roe, G. In defense of Milankovitch. *Geophys. Res. Lett.* **33**, L24703 (2006).

9. Ganopolski, A. & Brovkin, V. Simulation of climate, ice sheets and CO_2 evolution during the last four glacial cycles with an Earth system model of intermediate complexity. *Clim. Past* **13**, 1695–1716 (2017).

10. Buizert, C. et al. Greenland-wide seasonal temperatures during the last deglaciation. *Geophys. Res. Lett.* **45**, 1905–1914 (2018).

11. Petit, J. R. et al. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).

12. EPICA Community Members. Eight glacial cycles from an Antarctic ice core. *Nature* **429**, 623–628 (2004).

13. Bender, M. L. Orbital tuning chronology for the Vostok climate record supported by trapped gas composition. *Earth Planet. Sci. Lett.* **204**, 275–289 (2002).

14. Alley, R. B. & Anandakrishnan, S. Variations in melt-layer frequency in the GISP2 ice core: implications for Holocene summer temperatures in central Greenland. *Ann. Glaciol.* **21**, 64–70 (1995).

15. deMenocal, P. et al. Abrupt onset and termination of the African Humid Period: rapid climate responses to gradual insolation forcing. *Quat. Sci. Rev.* **19**, 347–361 (2000).

16. Huybers, P. Antarctica's orbital beat. *Science* **325**, 1085–1086 (2009).

17. Laepple, T., Werner, M. & Lohmann, G. Synchronicity of Antarctic temperatures and local solar insolation on orbital timescales. *Nature* **471**, 91–94 (2011).

18. Jones, T. R. et al. Southern Hemisphere climate variability forced by Northern Hemisphere ice-sheet topography. *Nature* **554**, 351–355 (2018).

19. White, J. W. C. et al. *Stable Isotopes of Ice in the Transition and Glacial Sections of the WAIS Divide Deep Ice Core* (USAP, 2019); https://doi.org/10.15784/601274.

20. Jones, T. R. et al. *Mid-Holocene High-Resolution Water Isotope Time Series for the WAIS Divide Ice Core* (USAP, 2020); https://doi.org/10.15784/601326.

21. Jones, T. R. et al. Improved methodologies for continuous-flow analysis of stable water isotopes in ice cores. *Atmos. Meas. Tech.* **10**, 617–632 (2017).

22. Sigl, M. et al. The WAIS Divide deep ice core WD2014 chronology— Part 2: annual-layer counting (0–31 ka bp). *Clim. Past* **12**, 769–786 (2016).

23. Vinther, B. M. et al. Climatic signals in multiple highly resolved stable isotope records from Greenland. *Quat. Sci. Rev.* **29**, 522–538 (2010).

24. Hughes, A. G. et al. High-frequency climate variability in the Holocene from a coastal-dome ice core in east-central Greenland. *Clim. Past* **16**, 1369–1386 (2020).

25. Morgan, V. & van Ommen, T. D. Seasonality in late-Holocene climate from ice-core records. *Holocene* **7**, 351–354 (1997).

26. Jones, T. R. et al. Water isotope diffusion in the WAIS Divide ice core during the Holocene and last glacial. *J. Geophys. Res. Earth Surf.* **122**, 290–309 (2017).

27. Johnsen, S. J. et al. in *Physics of Ice Core Records* (ed. Hondoh, T.) 121–140 (Hokkaido Univ. Press, 2000).

28. Vinther, B. M., Johnsen, S. J., Andersen, K. K., Clausen, H. B. & Hansen, A. W. NAO signal recorded in the stable isotopes of Greenland ice cores. *Geophys. Res. Lett.* **30**, 1387 (2003).

29. Markle, B. R. & Steig, E. J. Improving temperature reconstructions from ice-core water-isotope records. *Clim. Past* **18**, 1321–1368 (2022).

30. Cuffey, K. M. et al. Deglacial temperature history of West Antarctica. *Proc. Natl Acad. Sci. USA* **113**, 14249–14254 (2016).

31. Steig, E. J. et al. Recent climate and ice-sheet changes in West Antarctica compared with the past 2000 years. *Nat. Geosci.* **6**, 372–375 (2013).

32. Stenni, B. et al. Antarctic climate variability on regional and continental scales over the last 2000 years. *Clim. Past* **13**, 1609–1634 (2017).

33. Marcott, S. A., Shakun, J. D., Clark, P. U. & Mix, A. C. A reconstruction of regional and global temperature for the past 11,300 years. *Science* **339**, 1198–1201 (2013).

34. Kaufman, D. et al. Holocene global mean surface temperature, a multi-method reconstruction approach. *Sci. Data* **7**, 201 (2020).

35. Singarayer, J. S. & Valdes, P. J. High-latitude climate sensitivity to ice-sheet forcing over the last 120kyr. *Quat. Sci. Rev.* **29**, 43–55 (2010).

36. Shakun, J. D. et al. Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. *Nature* **484**, 49–54 (2012).

37. Marsicek, J., Shuman, B. N., Bartlein, P. J., Shafer, S. L. & Brewer, S. Reconciling divergent trends and millennial variations in Holocene temperatures. *Nature* **554**, 92–96 (2018).

38. Osman, M. B. et al. Globally resolved surface temperatures since the Last Glacial Maximum. *Nature* **599**, 239–244 (2021).

39. Bova, S., Rosenthal, Y., Liu, Z., Godad, S. P. & Yan, M. Seasonal origin of the thermal maxima at the Holocene and the last interglacial. *Nature* **589**, 548–553 (2021).

40. Liu, Z. et al. The Holocene temperature conundrum. *Proc. Natl Acad. Sci. USA* **111**, E3501–E3505 (2014).

41. Bader, J. et al. Global temperature modes shed light on the Holocene temperature conundrum. *Nat. Commun.* **11**, 4726 (2020).

42. Huybers, P. Combined obliquity and precession pacing of late Pleistocene deglaciations. *Nature* **480**, 229–232 (2011).

43. Briggs, R. D., Pollard, D. & Tarasov, L. A data-constrained large ensemble analysis of Antarctic evolution since the Eemian. *Quat. Sci. Rev.* **103**, 91–115 (2014).

44. Argus, D. F., Peltier, W. R., Drummond, R. & Moore, A. W. The Antarctica component of postglacial rebound model ICE-6G_C (VM5a) based on GPS positioning, exposure age dating of ice thicknesses, and relative sea level histories. *Geophys. J. Int.* **198**, 537–563 (2014).

45. Masson-Delmotte, V. et al. A review of Antarctic surface snow isotopic composition: observations, atmospheric circulation and isotopic modeling. *J. Clim.* **21**, 3359–3387 (2008).

46. Fortuin, J. P. F. & Oerlemans, J. Parameterization of the annual surface temperature and mass balance of Antarctica. *Ann. Glaciol.* **14**, 78–84 (1990).

47. Bandt, C. & Pompe, B. Permutation entropy: a natural complexity measure for time series. *Phys. Rev. Lett.* **88**, 174102 (2002).

48. Garland, J. et al. Anomaly detection in paleoclimate records using permutation entropy. *Entropy* **20**, 931 (2018).

49. Johnsen, S. J. Stable isotope homogenization of polar firn and ice. In *Proc. Symposium on Isotopes and Impurities in Snow and Ice* 210–219 (International Association of Hydrological Sciences, 1977).

50. Whillans, I. M. & Grootes, P. M. Isotopic diffusion in cold snow and firn. *J. Geophys. Res.* **90**, 3910–3918 (1985).

51. Cuffey, K. M. & Steig, E. J. Isotopic diffusion in polar firn: implications for interpretation of seasonal climate parameters in ice-core records, with emphasis on central Greenland. *J. Glaciol.* **44**, 273–284 (1998).

52. Gkinis, V., Simonsen, S. B., Buchardt, S. L., White, J. W. C. & Vinther, B. M. Water isotope diffusion rates from the NorthGRIP ice core for the last 16,000 years—glaciological and paleoclimatic implications. *Earth Planet. Sci. Lett.* **405**, 132–141 (2014).

53. Ciais, P. & Jouzel, J. Deuterium and oxygen 18 in precipitation: isotopic model, including mixed cloud processes. *J. Geophys. Res. Atmos.* **99**, 16793–16803 (1994).

54. Kavanaugh, J. L. & Cuffey, K. M. Space and time variation of $\delta^{18}O$ and δD in Antarctic precipitation revisited. *Glob. Biogeochem. Cycles* **17**, 1017 (2003).

55. Buizert, C. et al. Antarctic surface temperature and elevation during the Last Glacial Maximum. *Science* **372**, 1097–1101 (2021).

56. Stevens, C. M. et al. The Community Firn Model (CFM) v1. 0. *Geosci. Model Dev.* **13**, 4355–4377 (2020).

57. Gkinis, V. et al. Numerical experiments on firn isotope diffusion with the Community Firn Model. *J. Glaciol.* **67**, 450–472 (2021).

58. Agosta, C. et al. Estimation of the Antarctic surface mass balance using the regional climate model MAR (1979–2015) and identification of dominant processes. *Cryosphere* **13**, 281–296 (2019).

59. Arienzo, M. M. et al. Holocene black carbon in Antarctica paralleled Southern Hemisphere climate. *J. Geophys. Res. Atmos.* **122**, 6713–6728 (2017).

60. Casado, M., Münch, T. & Laepple, T. Climatic information archived in ice cores: impact of intermittency and diffusion on the recorded isotopic signal in Antarctica. *Clim. Past* **16**, 1581–1598 (2020).

61. Hwang, Y. T. & Frierson, D. M. W. Increasing atmospheric poleward energy transport with global warming. *Geophys. Res. Lett.* **38**, L09801 (2011).

62. Roe, G. H., Feldl, N., Armour, K. C., Hwang, Y. T. & Frierson, D. M. The remote impacts of climate feedbacks on regional climate predictability. *Nat. Geosci.* **8**, 135–139 (2015).

63. Siler, N., Roe, G. H. & Armour, K. C. Insights into the zonal-mean response of the hydrologic cycle to global warming from a diffusive energy balance model. *J. Clim.* **31**, 7481–7493 (2018).

64. Gordon, C. et al. The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Clim. Dynam.* **16**, 147–168 (2000).

65. Valdes, P. J. et al. The BRIDGE HadCM3 family of climate models: HadCM3@Bristol v1.0. *Geosci. Model Dev.* **10**, 3715–3743 (2017).

66. DiNezio, P. N. & Tierney, J. E. The effect of sea level on glacial Indo-Pacific climate. *Nat. Geosci.* **6**, 485–491 (2013).

67. Berger, A. & Loutre, M. F. Insolation values for the climate of the last 10 million years. *Quat. Sci. Rev.* **10**, 297–317 (1991).

68. Spahni, R. et al. Atmospheric methane and nitrous oxide of the late Pleistocene from Antarctic ice cores. *Science* **310**, 1317–1321 (2005).

69. Loulergue, L. et al. Orbital and millennial-scale features of atmospheric CH4 over the past 800,000 years. *Nature* **453**, 383–386 (2008).

70. Peltier, W. R., Argus, D. F. & Drummond, R. Space geodesy constrains ice age terminal deglaciation: the global ICE-6G_C (VM5a) model. *J. Geophys. Res. Solid Earth* **120**, 450–487 (2015).

71. Tarasov, L. & Richard Peltier, W. Greenland glacial history and local geodynamic consequences. *Geophys. J. Int.* **150**, 198–229 (2002).

72. Tarasov, L., Dyke, A. S., Neal, R. M. & Peltier, W. R. A data-calibrated distribution of deglacial chronologies for the North American ice complex from glaciological modeling. *Earth Planet. Sci. Lett.* **315**, 30–40 (2012).

73. Tarasov, L. et al. The global GLAC-1c deglaciation chronology, melwater pulse 1-a, and a question of missing ice. In *IGS Symposium: Contribution of Glaciers and Ice Sheets to Sea-Level Change* (International Geosynthetic Society, 2014).

74. Abe-Ouchi, A. et al. Insolation-driven 100,000-year glacial cycles and hysteresis of ice-sheet volume. *Nature* **500**, 190–193 (2013).

75. Steig, E. J. et al. Influence of West Antarctic Ice Sheet collapse on Antarctic surface climate. *Geophys. Res. Lett.* **42**, 4862–4868 (2015).

76. Fudge, T. J. et al. Variable relationship between accumulation and temperature in West Antarctica for the past 31 000 years. *Geophys. Res. Lett.* **43**, 3795–3803 (2016).

Acknowledgements

This work was supported by US National Science Foundation (NSF) grant nos. 0537593, 0537661, 0537930, 0539232, 1043092, 1043167, 1043518, 1142166 and 1807478. Field and logistical activities were managed by the WAIS Divide Science Coordination Office at the Desert Research Institute, Reno, NV, USA, and the University of New Hampshire, USA (NSF grant nos. 0230396, 0440817, 0944266 and 0944348). The NSF Division of Polar Programs funded the Ice Drilling Program Office, the Ice Drilling Design and Operations group, the National Ice Core Laboratory, the Antarctic Support Contractor and the 109th New York Air National Guard. K.M.C. was supported by The Martin Family Foundation. W.H.G.R. was funded by a Leverhulme Trust Research Project Grant. HadCM3 simulations were carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol (http://www.bris.ac.uk/acrc/) and Oswald, University of Northumbria. M.S. was supported by the European Research Council Grant under the European Union's Horizon 2020 research and innovation program (820047). We thank C. Agosta for providing the MAR outputs. We thank the Stable Isotope Lab at INSTAAR for their collective expertise in helping to measure the water-isotope datasets used in this study.

Author information

Authors and Affiliations

1. Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, USA

   Tyler R. Jones, Bradley R. Markle, Abigail G. Hughes, Valerie Morris, Bruce H. Vaughn, Kevin S. Rozmiarek & Chloe A. Brashear

2. Department of Geography, University of California, Berkeley, CA, USA

   Kurt M. Cuffey

3. Geography and Environmental Sciences, Northumbria University, Newcastle-upon-Tyne, UK

William H. G. Roberts

4. Department of Geological Sciences, University of Colorado, Boulder, CO, USA

Bradley R. Markle, Abigail G. Hughes, Kevin S. Rozmiarek & Chloe A. Brashear

5. Department of Earth and Space Sciences, University of Washington, Seattle, WA, USA

Eric J. Steig & T. J. Fudge

6. Cryospheric Science Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA

C. Max Stevens

7. Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA

C. Max Stevens

8. School of Geographical Sciences, University of Bristol, Bristol, UK

Paul J. Valdes

9. Climate and Environmental Physics, Physics Institute & Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

Michael Sigl

10. Center on Narrative, Disinformation and Strategic Influence, Arizona State University, Tempe, AZ, USA

Joshua Garland

11. Centre for Ice and Climate, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

Bo M. Vinther

12. College of Arts and Sciences, University of North Carolina, Chapel Hill, NC, USA

James W. C. White

Contributions

T.R.J. designed the project. T.R.J., K.M.C. and W.H.G.R. led the writing of the paper, with help from B.R.M. and E.J.S. High-resolution water-isotope measurements were contributed by T.R.J. and J.W.C.W. The analysis of the water isotopes was led by V.M., B.H.V., T.R.J. and J.W.C.W. HadCM3 simulations were conducted by W.H.G.R. and P.J.V. W.H.G.R. developed the methodology for determining how modelled sea ice affects West Antarctic temperature. B.R.M. conducted MEBM simulations. C.M.S. and T.R.J. conducted CFM simulations. T.R.J., K.M.C. and B.M.V. developed the diffusion-correction calculations, with help from A.G.H. and C.A.B. T.R.J. developed the methodology for the selection of extrema (summer and winter) in the diffusion-corrected water-isotope data. T.R.J. developed the methodology for quantifying the effect of seasonal accumulation on water-isotope diffusion using the CFM and impurity data. T.R.J. and K.M.C. determined the methodology for the uncertainty of seasonal temperature reconstruction. K.M.C. and E.J.S. determined the methodology for the uncertainty of multimillennial temperature trends, with help from T.R.J. B.R.M. provided isotope-temperature scaling uncertainty using SWIM. K.M.C. determined the methodology for constraints on WAIS Divide elevation changes in the Holocene. T.J.F. and M.S. provided impurity data. T.R.J., V.M., B.H.V., J.G. and K.S.R. assisted with the development of the water-isotope dataset over depths of 1,035.4–1,368.2 m. All authors discussed the results and contributed input to the manuscript.

Corresponding author

Correspondence to Tyler R. Jones.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Nicholas Golledge, Nerilie Abram and Mathieu Casado for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 WDC water-isotope data.

a, The raw, high-resolution WDC δD water isotope record[18,19,20] (grey), the raw 50-yr running mean (white), and the diffusion-corrected signal (black). **b**, The WDC diffusion-corrected δD record with extrema picks for summer (red) and winter (blue). **c**, The high-resolution diffusion length record (black; 140-yr windows, 70-yr time steps; 1σ uncertainty bounds in light grey) compared to prior estimates[26] (red; 500-yr windows, 500-yr time steps).

Source data

Extended Data Fig. 2 Precipitation uncertainties.

Uncertainties for seasonally weighted accumulation are shown in a-f, and for precipitation intermittency in g-h. **a**, The diffusion envelope of CFM output data (50-yr avg.), based on an input sine wave with $f = 1\text{yr}^{-1}$ and amplitude = 15.43‰. The original (input) amplitude of the signal (black, dotted-dashed lines) decreases as time passes due to diffusion and downward advection of the firn, as shown by the decay of the maximum (red) and minimum (blue) lines, while the mean values of the 'constant' and 'cycle' (black solid and dashed lines, respectively) scenarios do not change and are dependent on the seasonal weighting of snowfall. **b**, Diffused CFM output data from beneath the pore close-off depth (>200 yr) (black lines; smaller amplitude), with diffusion-corrected data shown with grey lines (larger amplitude). Red circles are the annual maximum value, and blue circles the annual minimum, selected using the same algorithm as Fig. 1a. **c**, A zoom of black carbon (BC) concentrations at ~6.5 ka[22,59]. The maxima (red circles) and minima (blue circles) can be used to separate approximate depth intervals corresponding to winter (BC_1) and summer (BC_2); vertical blue lines correspond to nominal January 1, as defined by the peak of nssS/Na[22]. **d**, The 140-yr averages for BC_1 (blue) and BC_2 (red). The grey line is WDC annual accumulation[76], orange circles are $BC_1 + BC_2$, which should equal annual accumulation. **e**, Black carbon seasonality BC_1/BC_2 (black), based on (d). **f**, Accumulation seasonality for HadCM3 seasonal snowfall (red line) compared to the range of seasonality tested using the CFM (dashed blue lines) and modern MAR seasonality[58] (blue diamond). **g**, Distribution of annual amplitudes for water isotopes for a 1,000-year window centred at 4 ka. **h**, Standard errors are determined for 1,000-realizations of random sampling of the distribution in (g) to determine a standard deviation of the residuals of the true mean minus the mean of the random sampling.

Extended Data Fig. 3 Trend analysis.

a,**b**, The first 10 of 10,000 randomly generated, alternative seasonal temperature histories for summer (a) and winter (b), used in Fig. 2b,d to generate probability distributions of temperature trends in the Holocene. Thick, solid lines are mean values, and thick, dashed lines are 2σ uncertainty ranges. **c**–**f**, The 1,000-year (thin line) and 300-year (thick line) averages normalized to the 11-0 ka mean (c–e) and residuals (1,000 year minus 300

year) (f) of summer (red), winter (blue), and the mean (black), used to calculate R^2 values in Extended Data Table 2.

Extended Data Fig. 4 MEBM results.

MEBM seasonal surface temperatures are shown in a–c. **a**, Modelled seasonal surface temperature cycle at 80°S, coloured by age. **b**,**c**, Zoom in of modelled summer and winter temperature. MEBM annual results for the mean, maximum, and minimum are shown in d–l. Plots for temperature anomaly normalized to the mean (blue), insolation (red), and heat divergence (black) for the annual mean (d,g,j), annual max (e,h,k), and annual min (f,i,l). Note the sign of heat divergence; negative values correspond to heat convergence at the site.

Extended Data Fig. 5 80°S energy balance at 10 ka.

Bar charts of HadCM3 energy balance terms at 10 ka for **a**, December (summer) and **b**, July (winter), including model runs for 'orbit only' (purple, ORBIT), 'ice sheet only' (blue, GLAC1D; green, ICE-6G), and 'all forcings' (orange, GLAC1D; yellow, ICE-6G). Positive values all indicate a surface or atmospheric warming. Variables include surface temperature (T_{surf} in °K), latent heat (LH in Wm^{-2}), sensible heat to the surface (SH_d in Wm^{-2}), shortwave radiation (SW in Wm^{-2}), downward LW radiation (LW_d in Wm^{-2}), and change in heat transport ($\nabla \cdot \mathbf{F}$ in 10^7 W).

Extended Data Fig. 6 Sea ice variability and temperature in HadCM3 simulations.

Maps of the dominant pattern of variability in each of the seasons and scatter plots of amplitude of the pattern against temperature. Maps were created using Python's package cartopy. **a**,**c**, The EOF of sea ice variability in the Southern Hemisphere for December and July in ALL of the simulations: this is the dominant pattern of sea ice variability. **b**,**d**, The amplitude of the patterns in (a) and (c) vs. the temperature at 80°S for each simulation. Plots e–h show the zonal-mean temperature and sea ice in HadCM3 simulations for December average. **e**, Change in sea ice fraction for coupled-model

simulations from 10 ka to 0 ka. **f**, Change in surface temperature between 10 ka and 0 ka for the coupled-model simulations (dotted line) and atmosphere-only simulations (solid line). **g**, Change in surface temperature for atmosphere-only runs from 10 ka to 0 ka. **h**, Change in surface temperature for atmosphere-only runs from 10 ka to 0 ka. See Extended Data Table 3 for descriptions of the model simulations used in panels (g) and (h).

Extended Data Fig. 7 Model results vs. WDC temperatures.

Coefficient of determination and p-values for comparison of HadCM3 (1-kyr resolution, n = 12) or MEBM (0.5-kyr resolution, n = 23) model results with WDC summer and winter temperatures.

Extended Data Table 1 CFM simulation inputs
Extended Data Table 2 Seasonal vs. mean correlations
Extended Data Table 3 Atmosphere-only climate model experiments

Source data

Source Data Fig. 1

Source Data Fig. 2

Source Data Fig. 3

Source Data Fig. 4

Source Data Extended Data Fig. 1

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The

images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Jones, T.R., Cuffey, K.M., Roberts, W.H.G. *et al.* Seasonal temperatures in West Antarctica during the Holocene. *Nature* **613**, 292–297 (2023). https://doi.org/10.1038/s41586-022-05411-8

- Received: 26 May 2021
- Accepted: 04 October 2022
- Published: 11 January 2023
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05411-8

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05411-8

- Article
- Open Access
- Published: 11 January 2023

Periclase deforms more slowly than bridgmanite under mantle conditions

- Patrick Cordier ,
- Karine Gouriet,
- Timmo Weidner,
- James Van Orman,
- Olivier Castelnau,
- Jennifer M. Jackson &
- …
- Philippe Carrez

Show authors

Nature **volume 613**, pages 303–307 (2023)

- 828 Accesses

- 62 Altmetric

- Metrics details

Abstract

Transport of heat from the interior of the Earth drives convection in the mantle, which involves the deformation of solid rocks over billions of years. The lower mantle of the Earth is mostly composed of iron-bearing

bridgmanite $MgSiO_3$ and approximately 25% volume periclase MgO (also with some iron). It is commonly accepted that ferropericlase is weaker than bridgmanite[1]. Considerable progress has been made in recent years to study assemblages representative of the lower mantle under the relevant pressure and temperature conditions[2,3]. However, the natural strain rates are 8 to 10 orders of magnitude lower than in the laboratory, and are still inaccessible to us. Once the deformation mechanisms of rocks and their constituent minerals have been identified, it is possible to overcome this limitation thanks to multiscale numerical modelling, and to determine rheological properties for inaccessible strain rates. In this work we use 2.5-dimensional dislocation dynamics to model the low-stress creep of MgO periclase at lower mantle pressures and temperatures. We show that periclase deforms very slowly under these conditions, in particular, much more slowly than bridgmanite deforming by pure climb creep. This is due to slow diffusion of oxygen in periclase under pressure. In the assemblage, this secondary phase hardly participates in the deformation, so that the rheology of the lower mantle is very well described by that of bridgmanite. Our results show that drastic changes in deformation mechanisms can occur as a function of the strain rate.

Main

This work aims to model steady-state creep of periclase (MgO) under lower mantle conditions, with an emphasis on the role of extremely low strain rates. It is based on the extensive knowledge we have of the deformation of this oxide following (1) the large number of experimental studies carried out on this ceramic since the 1970s, including the more recent mineral physics studies taking into account the effect of pressure, and (2) the important modelling effort carried out in recent years. At room pressure, periclase is easily deformed by glide of $\frac{1}{2} < 110 >$ dislocations on {110} planes. The thermal activation of this mechanism can now be modelled from the atomic scale[4], even at low strain rates[5]. Pressure strongly affects this mechanism and induces a transition of glide towards the {100} planes above 50 GPa[6,7,8,9]. At high temperature, dislocations glide without friction (athermal regime) and the flow stress results from dislocation interactions, leading, at large strains and strain rates, to a pronounced hardening[10]. At low

strain rates, ionic diffusion can enhance recovery mechanisms and allow steady-state creep[11]. In this study, we model the influence of mantle pressures on steady-state creep of periclase by 2.5-dimensional (2.5D) dislocation dynamics (DD). We focus on four pressures: 30, 60, 90 and 120 GPa, which, along the geotherm[12], correspond to 2,000, 2,300, 2,500 and 2,800 K respectively, at around depths of 800, 1,500, 2,100 and 2,700 km respectively.

Diffusion coefficients

Dislocation climb in MgO is enabled by diffusion of Mg and O atoms, and the climb rate is controlled by the slower diffusing species. Diffusion in MgO has been studied for decades, and the mechanisms and rates of cation and anion diffusion are well established[13]. Mg and O diffuse using vacant sites on the cation and anion sublattices, respectively. Because the concentration of cation vacancies is much higher than that of anion vacancies, in both nominally pure synthetic and natural periclase samples, O diffusion is orders of magnitude slower than Mg diffusion, and is the rate-limiting step in dislocation climb. Whereas Mg diffusion is controlled by extrinsic vacancies, produced to charge-balance trivalent (and tetravalent) cation impurities, O diffusion is intrinsic, enabled by thermally produced Mg and O vacancy pairs (Schottky defects).

To calculate the oxygen self-diffusion coefficient, ${D}_{{\rm{Ox}}}^{{\rm{sd}}}$, over the full range of temperatures and pressures relevant to Earth’s mantle, we use the expression presented by Ita and Cohen[14], which is derived from their theoretical results on vacancy formation and migration,

$${\rm{l}}{\rm{n}}{D}_{{\rm{O}}{\rm{x}}}^{{\rm{s}}{\rm{d}}}(P,T)={\rm{l}}{\rm{n}}({a}^{2}\,{\nu }_{{\rm{a}}})+\frac{{S}_{0}+P{S}_{0}^{{\prime} }}{{k}_{{\rm{B}}}}-\frac{{E}_{0}+P{V}_{0}+{P}^{2}{V}_{0}^{{\prime} }}{{k}_{{\rm{B}}}T}$$
(1)

where P and T are in GPa and K respectively, a is the lattice constant and k_B is the Boltzmann constant. The other parameters are listed in Table 1. The enthalpies are obtained by inverting the Arrhenius equation of diffusion coefficient:

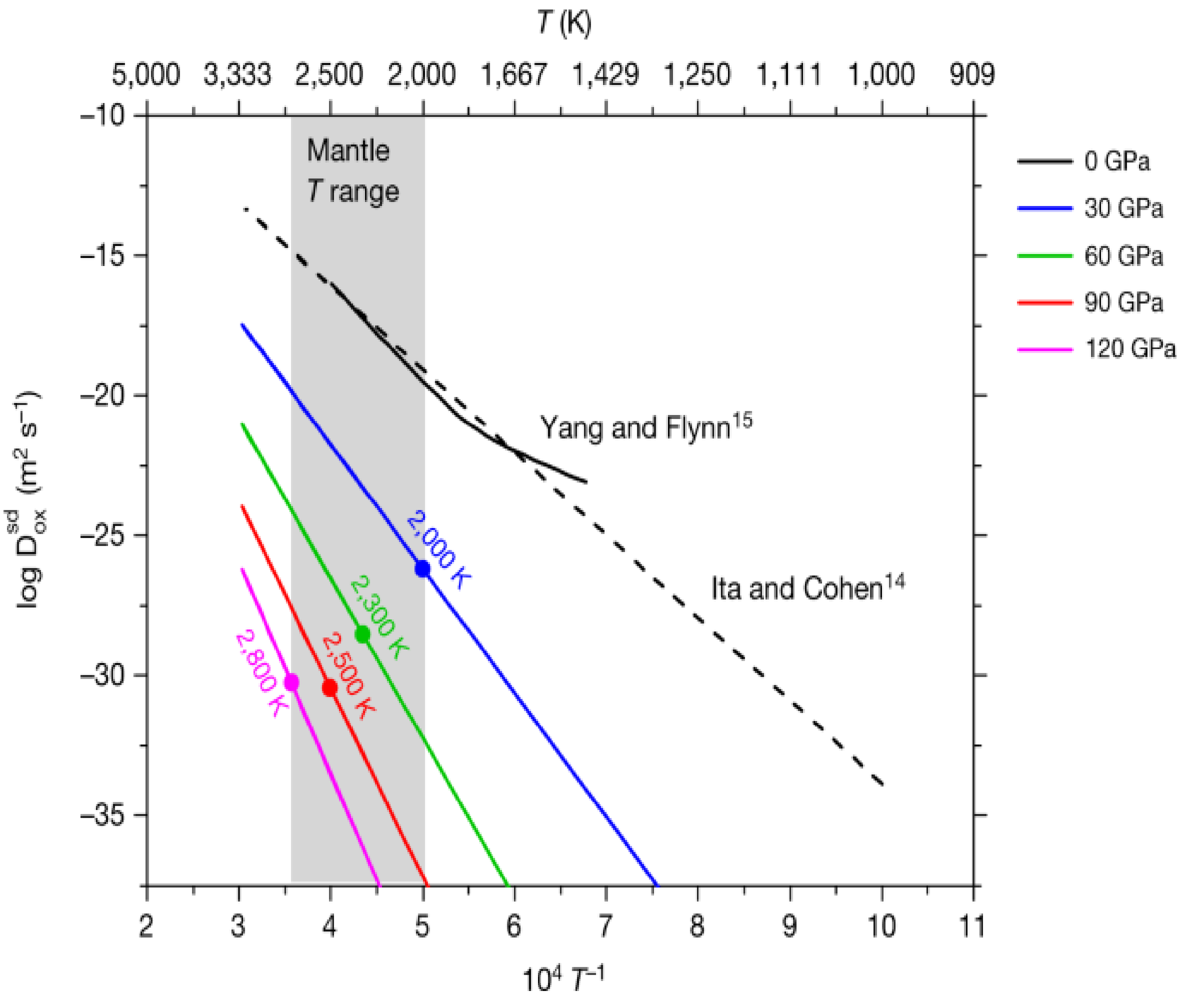
$${D}_{{\rm{O}}{\rm{x}}}^{{\rm{s}}{\rm{d}}}{={D}}_{0}\exp \left(\frac{-\Delta {H}^{{\rm{s}}{\rm{d}}}}{{k}_{{\rm{B}}}T}\right)$$
(2)

where $D_0 = 7.2 \times 10^{-4}$ m^2 s^{-1} and ΔH^{sd} are, respectively, the exponential prefactor and the activation enthalpy[11]. The activation enthalpies for each set of pressure/temperature and the diffusion coefficients are given in Table 2.

Table 1 Oxygen self-diffusion parameters for MgO global fit as a function of pressure and temperature (from Ita and Cohen[14])
Table 2 Oxygen self-diffusion coefficients (m^2 s^{-1}) from equation (1) and activation enthalpy (eV) for MgO obtained from equation (2)

This expression produces values of ${D}_{{\rm{Ox}}}^{{\rm{sd}}}$ that are in agreement with the most reliable experimental results on intrinsic O diffusion from Yang and Flynn[15] (Fig. 1). Higher values of ${D}_{{\rm{Ox}}}^{{\rm{sd}}}$ determined by Van Orman et al.[16] at high pressure are considered to reflect the contribution of dislocations[17] and/or other extended defects, as in most other studies of oxygen diffusion in MgO[13]. The results of Ita and Cohen[14] are also similar to those presented in several later theoretical studies[18,19,20].

Fig. 1: Oxygen diffusion coefficients in periclase.



The solid black line represents the diffusion coefficients of Yang and Flynn[15]. The black dashed line corresponds to the diffusion coefficients of Ita and Cohen[14] at 0 GPa. The other lines (blue, green, red and magenta) correspond to equation (1) at 30, 60, 90 and 120 GPa respectively. The range of temperatures considered in this study [2,000 K:2,800 K] is shown with the grey shading. The symbols (filled circles) correspond to the (*P*, *T*) conditions along the geotherm considered in the present study[12] (Table 2).

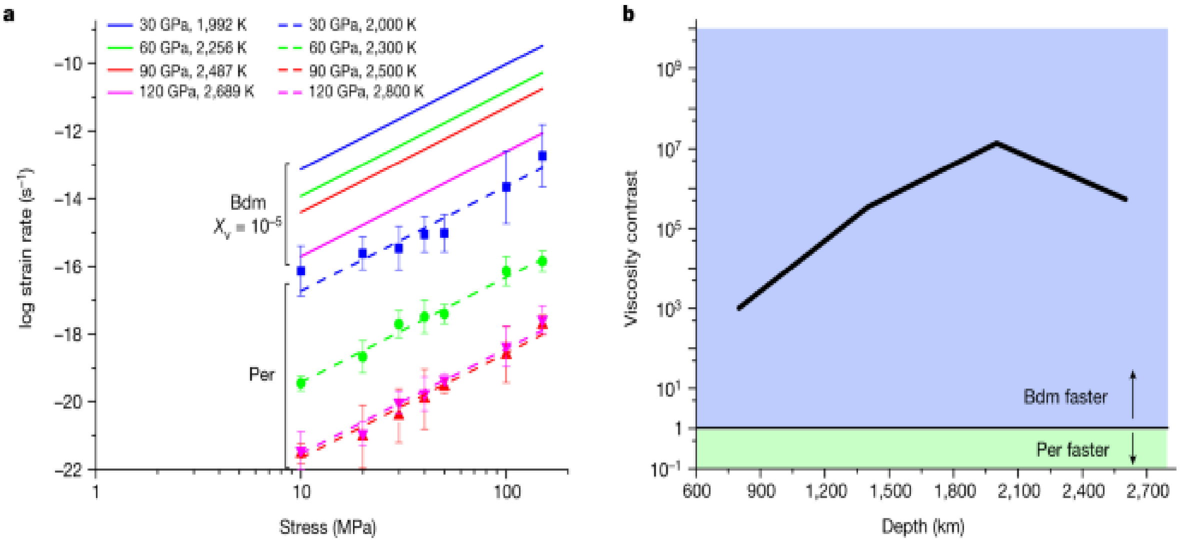
Source data

Modelling dislocation creep of periclase

To model plasticity of periclase under mantle conditions we use a 2.5D DD simulation approach already benchmarked for periclase at ambient

pressure[11]. DD simulation is a modelling tool that describes the collective motion and interactions of dislocations at the mesoscale. This technique is based on continuum elasticity theory, which provides the description of the elastic field induced by dislocations in a crystal, their interactions with each other and with respect to the stress field resulting from external loading. Within our DD code, dislocation glide is coupled with climb to investigate high-temperature creep (Methods). Our creep results are presented in Fig. 2a, which shows the creep strain rates obtained from the 2.5D DD simulations as a function of the creep stress for the four (P, T) conditions selected. A constant slope is found for all these strain rate–stress curves, which gives the value of the power law stress exponent, $n = 3.1$. A strong pressure dependence is found between (30 GPa, 2,000 K) and (90 GPa, 2,500 K), whereas creep rates at (90 GPa, 2,500 K) and (120 GPa, 2,800 K) are very close due to very similar values for the diffusion coefficients (Fig. 1 and Table 2). On Fig. 2a we compare the creep rates of periclase with those of bridgmanite at the same depths. For bridgmanite, we use the results of Reali et al.[21] for pure climb creep (PCC) with a vacancy concentration $X_v = 1 \times 10^{-5}$; these results were shown to be in good agreement with empirical geophysical constraints on the viscosity profile of Earth's lower mantle. Bridgmanite and periclase exhibit similar stress exponents. However, we see that these calculations lead to bridgmanite having higher creep rates than periclase at any depth. To better illustrate this difference, we report, on Fig. 2b, the viscosity contrast between those two phases, taken as the ratio \({\eta }_{{\rm{Per}}}/{\eta }_{{\rm{Bdm}}}\). We see that under lower mantle conditions bridgmanite creep rates are 10^3 to 10^7 greater than periclase creep rates. This result reflects the fact that in the lower mantle, anionic diffusion in periclase is much slower than cationic diffusion in bridgmanite (Extended Data Fig. 2). In the online Methods section and Extended Data Figs. 1–4 we consider other diffusion data from the literature for periclase and other vacancy concentrations in bridgmanite and show that the conclusions stated above remain valid in any case.

Fig. 2: Creep rate modelling.



a, Creep strain rates as a function of applied stress of bridgmanite (Bdm; solid line) and periclase (Per; symbols) at 30 GPa (blue), 60 GPa (green), 90 GPa (red) and 120 GPa (magenta). **b**, Viscosity contrast ${\eta }_{{\rm{Per}}}/{\eta }_{{\rm{Bdm}}}$ (calculated at 10 MPa) between the two pure phases, periclase and bridgmanite, as a function of depth.

Source data

Laboratory versus nature

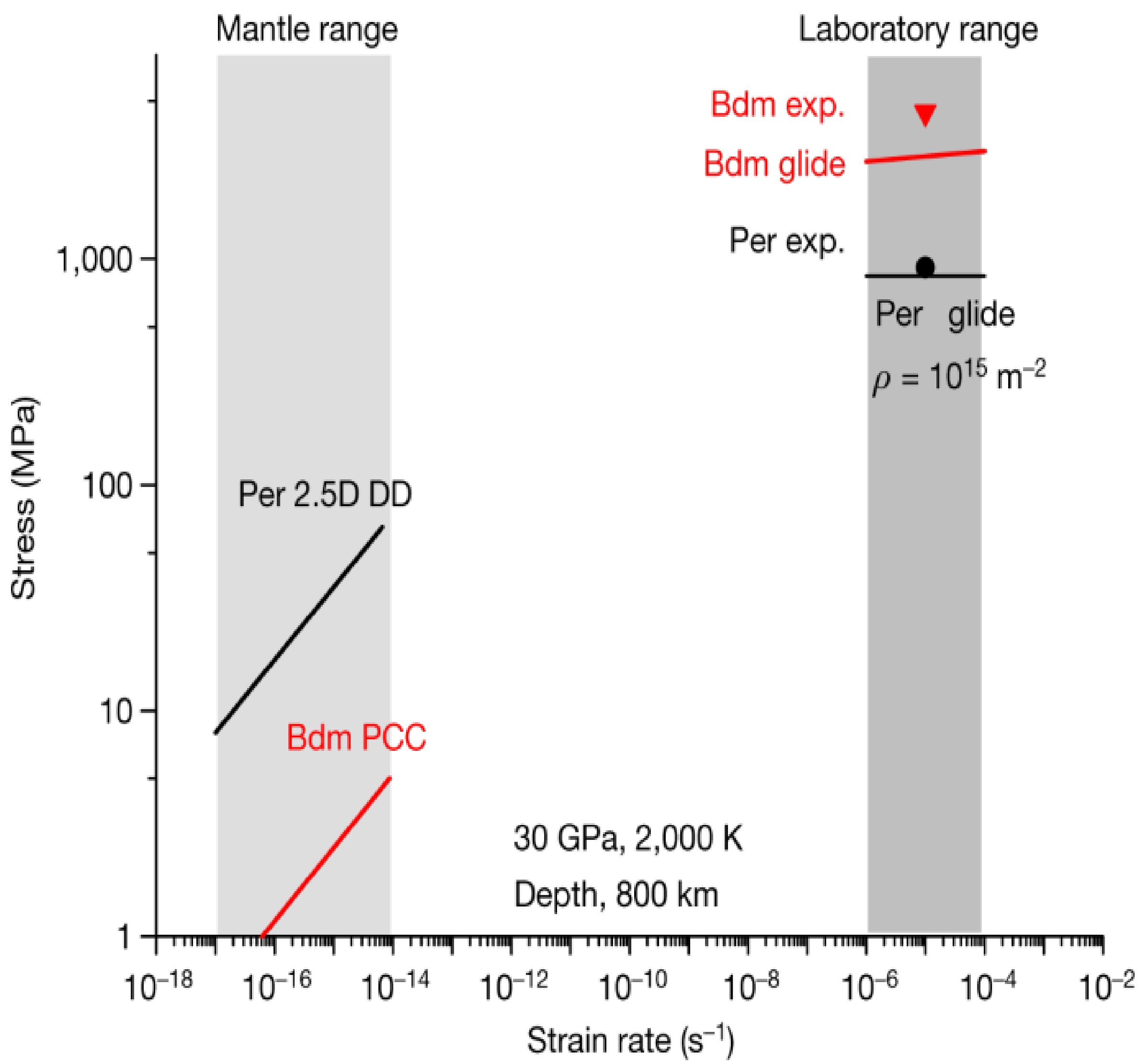
In this paragraph we address the influence of strain rate on deformation mechanisms. Our calculations showing that bridgmanite deforms faster than periclase under mantle conditions seem to contradict the common intuition that the silicate is stronger than the oxide. Beyond general considerations[1] or studies of analogous compounds[22,23], Girard et al.[3] conducted experiments under mantle conditions of pressure and temperature that show very clearly that ferropericlase accommodates most of the strain compared to bridgmanite. We show here that there is no fundamental opposition between modelling and experiments and that our models are able to account equally for natural and experimental conditions. The experiments of Girard et al.[3] are conducted at a strain rate of 3×10^{-5} s^{-1} and under very high stresses. Under these laboratory conditions, dislocation glide is expected as the dominant deformation mode. For bridgmanite, dislocation glide is thermally activated under these conditions and a model based on the Orowan equation

can be built[24]. The stress values predicted by this model are reported on Fig. 3. They are close to the experimental values (slightly lower as this model was constructed for single crystals and does not involve grain boundaries). At 30 GPa and 2,000 K, dislocations in periclase glide in the athermal regime[6]. Several deformation experiments of periclase report that in the athermal regime, no steady state is reached and that strain hardening occurs[25,26,27] under the influence of dislocation–dislocation interactions. The origin of this strain hardening has been studied and modelled by Amodeo et al.[10]. It is well described by the Taylor mechanism, and the athermal stress τ_f required to overcome the forest dislocation network is given by:

$${\tau }_{f}=\mu b\frac{{\rm{ln}}\left(1/b\sqrt{\beta {\rho }_{{\rm{f}}}}\right)}{{\rm{ln}}\left(1/b\sqrt{\beta {\rho }_{{\rm{ref}}}}\right)}\sqrt{\beta {\rho }_{{\rm{f}}}}$$
(3)

where μ is the shear modulus, b the modulus of the Burgers vector, β is a forest strengthening coefficient calibrated by Amodeo et al.[10], ${\rho }_{{\rm{f}}}$ is the forest dislocation density and ${\rho }_{{\rm{ref}}}$ is the reference dislocation density. Transmission electron microscopy observations[28] on the specimens of Girard et al.[3] confirm this view and show a very large dislocation density estimated at over $10^{15}\ m^{-2}$. Using this dislocation density and equation (3) leads to an excellent description of experimental values for the deformation of ferropericlase under shallow lower mantle P, T conditions at laboratory strain rates (Fig. 3). Our modelling approach allows us to reproduce very well the experimental data of Girard et al.[3] and confirm that, under laboratory conditions, bridgmanite is indeed expected to be stronger than ferropericlase. It should be noted that the flow stress value for ferropericlase corresponds to a very high hardening and that under more 'gentle' laboratory deformation conditions (that is, smaller strains or more recovery), the rheology contrast under laboratory conditions should be even larger. At 30 GPa, oxygen diffusion is already slow, and strain rates much lower than those obtainable in laboratory experiments would be required for climb to become active and promote the recovery necessary for steady-state creep.

Fig. 3: Influence of strain rate on deformation mechanisms.



This figure corresponds to approximately 30 GPa, 2,000 K, where the experimental (exp.) data of Girard et al.[3] (symbols) can be compared to numerical models (lines). At laboratory stresses and strain rates numerical models based on dislocation glide are in fair agreement with experimental data and show that bridgmanite (Bdm) is stronger than periclase (Per). At mantle strain rates, creep (Weertman creep modelled by 2.5D DD in periclase (see text) and pure climb creep (PCC) in bridgmanite) is controlled by ionic diffusion, which leads to an inversion of rheology, with periclase deforming more slowly than bridgmanite.

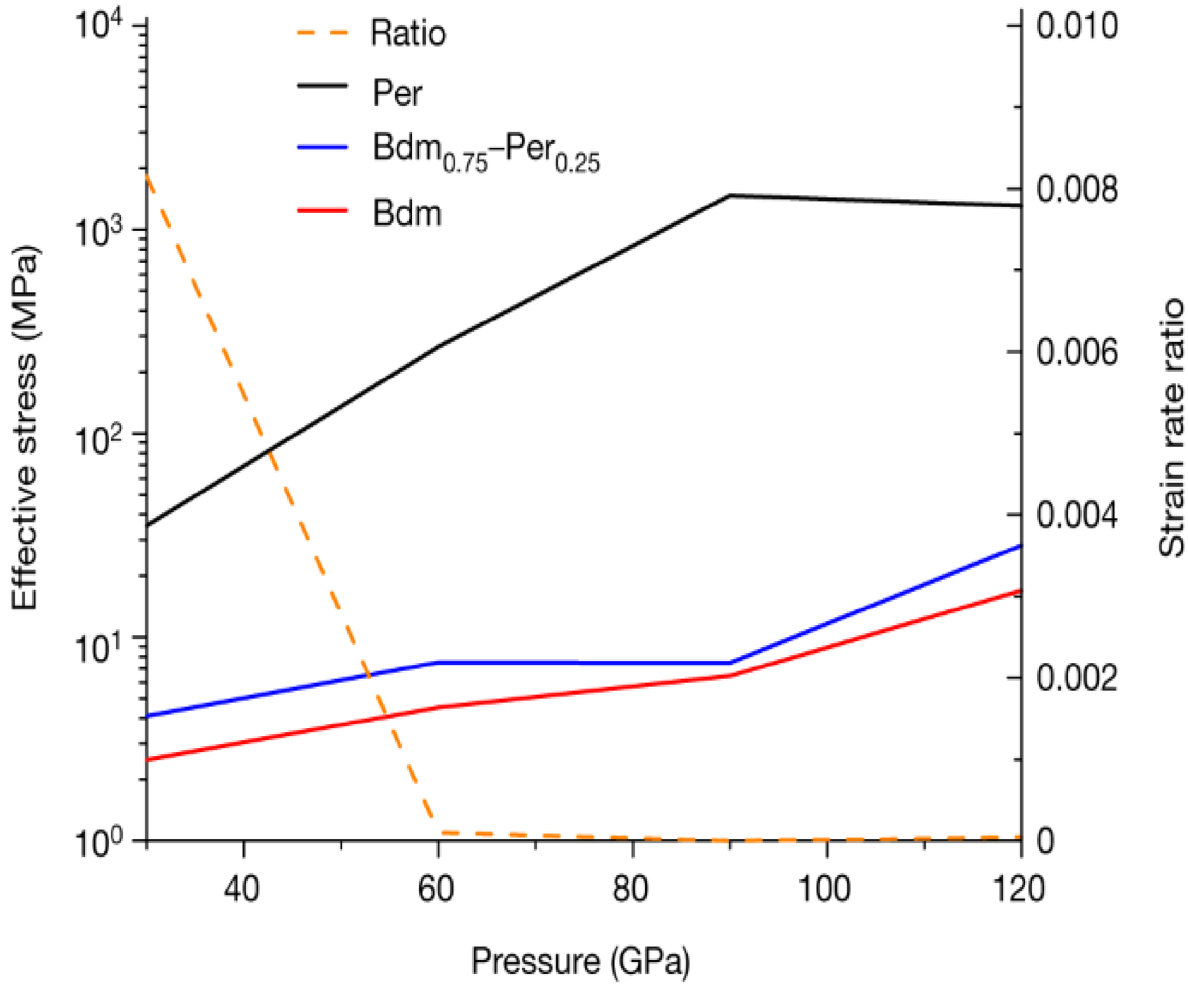
Source data

Deformation of a bridgmanite-periclase aggregate

Our model places first-order constraints on the intrinsic properties of the major phases of the lower mantle. The rheological behaviour of rocks in the lower mantle certainly involves more complexity. A first step towards this direction is to use dedicated micromechanical models, to connect the viscoplastic behaviour of a polycrystalline aggregate representative of a lower mantle assemblage with those of the individual phases, bridgmanite and periclase, and, at the same time, quantify the contribution of both phases to the effective strain rate, in relation with the aggregate microstructure.

We have therefore applied a state-of-the-art scale transition homogenization method[29] (partially optimized second-order self-consistent procedure, denoted POSO-SC) adapted for multiphase aggregates with nonlinear rheologies, such as those considered here. The model considers an isotropic microstructure, that is, a random mixture of bridgmanite and periclase with no preferred elongated shape nor crystallographic texture for each phase. Details are given in the Methods. The effective viscoplastic behaviour can be expressed with a nonlinear creep law with a stress sensitivity $n = 3.1$ (see Methods). Figure 4 shows the results obtained for a vacancy concentration $X_v = 1 \times 10^{-5}$ in bridgmanite and a volume fraction of periclase of 25% (alternative figures considering different combinations of diffusion coefficients are presented in Extended Data Fig. 5). For pressures in the range 30–120 GPa, the effective rheology (blue line) lies very close to that of pure bridgmanite (red line), with the flow stress values differing by a factor of only around 1.65, whereas the flow stress of pure periclase (black line) is one to two orders of magnitude larger. Periclase deforms at least 100 times slower than bridgmanite in the aggregate and its contribution to the effective strain rate is negligible. In the online Methods section, we show that the volume fraction of periclase would have to exceed 40% to significantly affect the rheology of the assemblage. The presence of iron in MgO may affect anionic diffusivity but does not change our conclusions for the bulk lower mantle (see online Methods section).

Fig. 4: Rheology of bridgmanite–periclase aggregates.



Pressure dependent flow stress \({\widetilde{\sigma }}_{0}\) for a bridgmanite (Bdm)–periclase (Per) aggregate deformed in creep at 10^{-15} s^{-1}. Calculations are performed within the POSO-SC scheme, with a volume fraction of periclase of 0.25 and a vacancy concentration in bridgmanite of $X_v = 1 \times 10^{-5}$. The ratio of strain rate (periclase/bridgmanite) is also indicated. Flow stresses σ_0 of pure bridgmanite (in red) and pure periclase (in black) are also shown for comparison.

Source data

Following the widespread acceptance that ferropericlase is the weakest phase[1], and inspired by the experimental results of Girard et al.[3], several studies[30,31,32] have discussed the implication of the morphology of ferropericlase (either forming a load-bearing framework or an interconnected weak layer[33]) on the rheology of the lower mantle. Our study

shows that under the conditions (mainly strain rate) of the lower mantle, the behaviour of periclase is radically different and that this fundamental assumption is now challenged. Periclase, minor in volume, hardly deforms and probably induces little evolution of the microstructure. Consequently, the rheology of the lower mantle is very well described by that of bridgmanite deforming by pure climb creep.

Methods

Dislocation dynamics calculations

The 2.5D DD simulation method is used throughout this study to model the creep properties of periclase under lower mantle conditions under a constant creep stress ranging from 10 to 150 MPa. We consider the *P*, *T* conditions on the geotherm of Stacey and Davis[12] at 30, 60, 90 and 120 GPa, and use the periclase elastic properties from Karki et al.[34] (Extended Data Table 2).

To capture the statistically representative volume element of the microstructure, dislocations of Burgers vector **b** = $\frac{1}{2}$<110> are introduced edge-on in two intersecting slip systems within a square element of typical cell size $L_x = L_y$ between 8 to 60 µm (function of the desired dislocation densities) and periodic boundary conditions are applied. As extensively described in some of our previous works[11], specific characteristics of a realistic MgO microstructure are introduced as local rules, including junction formation strengthening and a dislocation multiplication rate to allow random introduction of fresh dislocations of opposite signs on all slip systems. All local rules are parameterized according to experimental observations or three-dimensional DD simulations as described in Reali et al.[11].

For a realistic simulation of the creep behaviour, we use a combination of glide and climb events to control the displacement of dislocations in the simulation volume. The Burgers vector contained in the reference plane defines the slip direction, whereas the climb direction is defined as the one in the reference plane, orthogonal to the Burgers vector. The positive orientation of the climb direction is taken along the vacancy emission direction. During a given timestep of the simulation, dislocations are

displaced along the glide or climb direction according to the velocity formulation presented below.

In our simulations, dislocations glide in the athermal regime according to a « free-flight » velocity v_g, which is a function of the effective resolved shear stress $\tau = \tau_{app} + \tau_{int}$ (that is, including the applied resolved shear stress and the elastic interaction stress):

$${v}_{{\rm{g}}}=\frac{b\tau }{B}$$

As, in the range of *P* and *T* conditions investigated here, the lattice friction of MgO vanishes[5,6], the glide velocity is thus expressed with a viscous drag coefficient *B*. Such a coefficient is known to be temperature and material dependent. Thus, we consider a linear scaling of *B* with temperature[35] and use as a starting point the experimental value recorded at 300 K[36].

Climb events occur through the self-diffusion of vacancies adsorbed or emitted by the dislocation line. To describe the climb rate in steady-state creep equilibrium conditions, we use a dislocation climb velocity v_c expressed as follows[35]:

$${v}_{{\rm{c}}}=\alpha \frac{{D}^{{\rm{sd}}}}{b}\left[\exp \left(\frac{{\tau }_{{\rm{c}}}\,\varOmega }{{k}_{{\rm{B}}}T}\right)-\frac{{c}_{{\rm{\infty }}}}{{c}_{0}}\right]$$

where τ_c is the climb stress, calculated similarly to τ but resolved here along the climb direction. $\alpha =2\pi /{\rm{ln}}(R/{r}_{{\rm{c}}})$ is a geometrical factor that describes the cylindrical geometry of the vacancy flux field around the dislocation line, where *R* and ${r}_{{\rm{c}}}$ represent the two radii of the cylindrical surfaces through which the vacancy flux is calculated. Again, in the range of temperature investigated, the dislocation line must contain a high density of jogs, and thus be saturated with jogs, enabling vacancies to be absorbed or emitted instantaneously by the dislocation. Here, ${r}_{{\rm{c}}}$ is the dislocation core capture radius and *R* is taken as a fraction of the average dislocation distance. Being within the logarithmic term, the $R/{r}_{{\rm{c}}}$ ratio does not significantly affect the climb velocity values and here is taken as constant

and equal to 100[37]. Ω is the formation volume of vacancies in the Mg and O sites and is calculated from the unit cell volume of MgO—that is, $\varOmega ={a}^{3}/Z$ where $Z = 4$ is the number of formula units per unit cell. c_{∞} and c_0 are the vacancy concentrations far from the dislocation and the equilibrium vacancy concentration in the bulk volume, respectively. Following Boioli et al.[37], we assume that far away from the dislocations (that is, on the external limit R of the cylindrical flux) the vacancy concentration is constant and equal to the equilibrium concentration in the bulk volume ($c_{\infty} = c_0$). Finally, D_{sd} is the vacancy self-diffusion coefficient, which controls the flow of atomic species from and to the climbing dislocation.

Micromechanical model of the rheology of the aggregate

Local behaviour

We apply state-of-the-art formulations of nonlinear homogenization methods to estimate and bound the effective viscosity of a two-phase material constituted by bridgmanite and periclase with various volume fractions, at pressures and temperatures relevant for the lowermost part of Earth's mantle. These phases are supposed to be randomly mixed so that the effective mechanical behaviour of the whole aggregate can be considered isotropic. Bridgmanite and periclase are considered to be two isotropic viscous phases too. In reality, they are both made of grains, but we consider that these grains are equiaxed and randomly oriented so that, on average, the polycrystalline aggregates can be replaced by two homogenous and isotropic phases. The role of grain or phase boundaries is not explicitly taken into account. We are interested in the effective rheological behaviour in the permanent (that is, secondary) creep regime. Thus, the elastic behaviour of both phases does not come into play. At the scale of each phase, the viscoplastic behaviour can be described by the following nonlinear creep law:

$${\dot{\varepsilon }}_{{\rm{eq}}}={\dot{\varepsilon }}_{0}{\left(\frac{{\sigma }_{{\rm{eq}}}}{{\sigma }_{0}}\right)}^{n}$$

with ${\dot{\varepsilon }}_{0}$ a reference strain rate, σ_0 a reference stress, n the stress sensitivity, ${\dot{\varepsilon }}_{{\rm{eq}}}$ the equivalent strain rate and σ_{eq} the equivalent stress, defined as:

$${\dot{{\varepsilon }}}_{{\rm{e}}{\rm{q}}}=\sqrt{\frac{2}{3}{\dot{{\varepsilon }}}_{ij}{\dot{{\varepsilon }}}_{ij}},{\dot{{\sigma }}}_{{\rm{e}}{\rm{q}}}=\sqrt{\frac{3}{2}{{\sigma }{\prime} }_{ij}{{\sigma }{\prime} }_{ij}}$$

with summation over repeated indices (Einstein convention), ${\dot{\varepsilon }}_{{ij}}$ and ${\sigma {\prime} }_{{ij}}$ components of the strain rate tensor and deviatoric stress tensor, respectively. We chose here to take ${\dot{\varepsilon }}_{0}=1\times {10}^{-15}{{\rm{s}}}^{-1}$, that is, a value typical for lower mantle convection. Doing so, the σ_0 values are of the order of magnitude of the equivalent stress encountered in situ during mantle flow. Values of σ_0 for bridgmanite and periclase are fitted from the creep laws of Fig. 2 and available from the data deposit. The stress sensitivity n is the same for both phases, $n = 3.1$.

Effective behaviour

The effective behaviour reads in a similar way as the local behaviour:

$${\dot{\bar{\varepsilon }}}_{{\rm{eq}}}={\dot{\varepsilon }}_{0}{\left(\frac{{\bar{\sigma }}_{{\rm{eq}}}}{{\widetilde{\sigma }}_{0}}\right)}^{\widetilde{n}}$$

where $\dot{\bar{\varepsilon }}$ and $\bar{\sigma }$ are the effective strain rate and stress tensors, respectively, given by the volume average (denoted $\left\langle \bullet \right\rangle$) of local strain rate and stress fields: $\dot{\bar{\varepsilon }}=\left\langle \dot{\varepsilon }\right\rangle$ and $\bar{\sigma }=\left\langle \sigma \right\rangle$. $\widetilde{n}$ is the effective stress sensitivity, which can, in general, be inferred from the homogenization procedure. In the case considered here, things are simpler as bridgmanite and periclase exhibit the same n value. Therefore $\widetilde{n}=n=3.1$. Finally, ${\widetilde{\sigma }}_{0}$ is the

effective flow stress associated with the (nonlinear) viscosity of the aggregate in situ. It is determined by the homogenization procedure detailed below.

Homogenization procedure

To describe the possible rheology of an assemblage representative of the lower mantle, one can reasonably assume that bridgmanite and periclase phases are randomly mixed. In that case, a good estimation of the effective behaviour is provided by the Self-Consistent scheme (denoted SC below), based on a fully disordered microstructure[38,39]. To extend the SC scheme to nonlinear rheologies, as here, we have made use of the very efficient and accurate linearization procedure (so-called partially optimized second-order procedure (POSO)) of Ponte Castaneda[29].

Sensitivity to calculation parameters

The calculations presented in the main text are made based on the diffusion coefficients for pure MgO periclase and for $MgSiO_3$ bridgmanite with a vacancy concentration of 10^{-5}. For the calculation of the aggregate rheology, we considered a periclase proportion of 25% by volume. In the following, reasonable alternatives to these choices are discussed to assess the robustness of our conclusions.

Diffusion coefficients

As the creep rate critically depends on the diffusion coefficients, we have re-calculated the creep rates of periclase using another dataset of diffusion coefficients for oxygen in periclase, those of Yoo et al.[40], which are faster than those of Ita and Cohen[14] and those of Yang and Flynn[15]. Fitting equation (1) on the data of Yoo et al.[40] (Extended Data Fig. 1), we determine two new constants for equation (1): $E = 7200 \times 10^{-21}$ J and \({S}_{0}=2{k}_{{\rm{B}}}\). The activation enthalpies obtained with these new constants are given in Extended Data Table 1 and the evolution of the two sets of diffusion coefficients are compared in Extended Data Fig. 2. The creep rates of periclase with the diffusion coefficients of Yoo et al.[40] are

then slightly faster than with the diffusion coefficients of Ita and Cohen[14], but still significantly lower than those of bridgmanite (Extended Data Fig. 3). For this comparison, we have also taken into account the uncertainty on the rheology of bridgmanite owing to the lack of constraints on diffusion coefficients (this is discussed in detail by Reali et al.[21]). For this purpose, we vary the concentration of vacancies in bridgmanite X_v between 10^{-3} and 10^{-6}. Extended Data Fig. 2 shows that no matter which combination of data we may choose, Mg (or Si) diffusion in bridgmanite is always significantly faster than oxygen diffusion in periclase. The contrast is particularly marked at high pressures. Extended Data Figs. 3–7 show that even taking into account these uncertainties in the most conservative way does not change any of our conclusions.

Influence of iron

Our study is based on the end-member compositions MgO and $MgSiO_3$, whereas in the mantle these minerals contain iron (and aluminium for bridgmanite). These differences in composition are likely to affect the values of the diffusion coefficients on which our calculations are based. In bridgmanite, this is taken into account by the range of vacancy concentrations considered[21], see discussion above and Extended Data Figs. 3, 4, 6 and 7. In ferropericlase, it has been shown that oxygen diffusion coefficients are unaffected by the presence of trivalent cations[13,41], because these do not alter the concentration of neutral cation–anion vacancy pairs that are responsible for oxygen transport. The addition of significant amounts of FeO, however, is likely to alter oxygen diffusion coefficients. This effect was considered by Reali et al.[42], who showed that anion diffusion in MgO and in other oxides and halides with rock salt structures is well described by a homologous temperature scaling, wherein the diffusion coefficients as a function of temperature collapse onto a common curve when normalized to the melting temperature. The addition of FeO lowers the melting temperature of periclase, and hence is expected to enhance oxygen diffusivity. We have thus calculated oxygen diffusion coefficients under lower mantle conditions for ferropericlase as a function of its Fe content, under the same assumptions made by Reali et al.[42]. Considering the addition of up to 20 mol% FeO in periclase[43], we do find that the enhancement of

oxygen diffusion is up by a factor of approximately 2000 relative to pure MgO. However, this enhancement does not change the conclusions of this study for the bulk lower mantle, as oxygen diffusion in ferropericlase remains significantly slower than Mg or Si in bridgmanite.

Influence of periclase volume fraction

We have computed the rheology of bridgmanite–periclase aggregates for various volume fractions of periclase based on four homogenization procedures (Extended Data Figs. 6 and 7), the difference between them being related to the knowledge on the specimen microstructure that is taken into account.

First, Reuss and Voigt bounds are valid for any microstructures. They only depend on the volume fraction of both phases, not on their geometrical arrangement. They provide rigorous lower and upper bounds for ${\widetilde{\sigma }}_{0}$, respectively. In other words, ${\widetilde{\sigma }}_{0}$ in the aggregate cannot be smaller than ${{\widetilde{\sigma }}_{0}}^{{\rm{Reuss}}}$ nor larger than ${{\widetilde{\sigma }}_{0}}^{{\rm{Voigt}}}$. The Reuss bound is obtained when assuming that both phases are submitted to the same stress, whereas the Voigt bound is obtained when assuming a uniform strain rate in both phases.

The Hashin-Strikman variational upper bound[29], denoted HS+ in Extended Data Figs. 6 and 7, is another rigorous upper bound for ${\widetilde{\sigma }}_{0}$. It is more stringent than the Voigt upper bound as is based on an extra assumption concerning the specimen microstructure, that is, that the microstructure is isotropic. It is obtained when assuming that the softer phase (here, bridgmanite) is spread as inclusions in a stiffer matrix (periclase) while the associated nonlinear homogenization problem is addressed using the variational procedure introduced by Ponte Castaneda, also known as ‘modified secant’ method[44].

To stick more closely to the possible rheology of an assemblage representative of the lower mantle, one can reasonably assume that bridgmanite and periclase phases are randomly mixed, that is, that contrary

to what is assumed within the HS+ bound, both phases are on the same footing, with neither of them playing the role of a matrix and the other one being inclusions. In that case, a good estimation of the effective behaviour is provided by the POSO extension of the Self-Consistent scheme (SC in Extended Data Figs. 6 and 7), introduced in the previous section. Here, unlike previous applications to polycrystalline aggregates, where individual crystal orientations were taken into account[45,46], a two-phase aggregate was considered. A fully optimized linearization procedure for the SC scheme has been proposed recently[47,48] and applied to minerals of the mantle transition zone[49]; results are close to the POSO version when the mechanical contrast between the phases and the nonlinearity of their behaviour is not loo large, as here.

Typical results are shown in Extended Data Figs. 6 and 7, calculated here for pressures between 30 and 120 GPa and $X_v = 10^{-3}$–10^{-6}. Due to the large mechanical contrast between periclase and bridgmanite, the effective rheology of the aggregate, predicted by the POSO-SC model, lies very close to the Reuss lower bound for volume fractions lower than 30%, whereas it is close to the Voigt upper bound for a volume fraction larger than 70%. In between, the effective stress \({\widetilde{\sigma }}_{0}\) evolves significantly, by more than one order of magnitude. This would correspond to a change in viscosity of more than three orders of magnitude, the viscosity being proportional to \({\left({\widetilde{\sigma }}_{0}\right)}^{n}\). One other specific feature of these results is illustrated in the right panels of Extended Data Figs. 6 and 7, which show the mean equivalent strain rate in the two phases within the aggregate, normalized by the effective strain rate, as predicted by the SC model. It is obtained that the periclase strain rate is really small at volume fractions lower than 40%. This volume fraction corresponds to a ‘mechanical percolation threshold’ observed in previous work[50,51]. At small volume fractions of the highly viscous periclase phase, the aggregate can deform with almost no plastic strain in that phase.

Data availability

The data of this manuscript are available at https://doi.org/10.5281/zenodo.5970733. Source data are provided with this

paper.

Code availability

The source code of the 2.5D DD is available at https://doi.org/10.5281/zenodo.6306546.

References

1. Yamazaki, D. & Karato, S.-I. Some mineral physics constraints on the rheology and geothermal structure of Earth's lower mantle. *Am. Mineral.* **86**, 385–391 (2001).

2. Miyagi, L. & Wenk, H.-R. Texture development and slip systems in bridgmanite and bridgmanite + ferropericlase aggregates. *Phys. Chem. Minerals* **43**, 597–613 (2016).

3. Girard, J., Amulele, G., Farla, R., Mohiuddin, A. & Karato, S. I. Shear deformation of bridgmanite and magnesiowüstite aggregates at lower mantle conditions. *Science* **351**, 144–147 (2016).

4. Amodeo, J., Carrez, P. H., Devincre, B. & Cordier, P. Multiscale modelling of MgO plasticity. *Acta Mat.* **59**, 2291–2301 (2011).

5. Cordier, P., Amodeo, J. & Carrez, P. H. Modelling the rheology of MgO under Earth's mantle pressure, temperature and strain-rates. *Nature* **481**, 177–180 (2012).

6. Amodeo, J., Carrez, P. H. & Cordier, P. Modelling the effect of pressure on the critical shear stress of MgO single crystals. *Philos. Mag.* **92**, 1523–1541 (2012).

7. Girard, J., Chen, J. & Raterron, P. Deformation of periclase single crystals at high pressure and temperature: quantification of the effect of pressure on slip-system activities. *J. Appl. Phys.* **111**, 112607 (2012).

8. Marquardt, H. & Miyagi, L. Slab stagnation in the shallow lower mantle linked to an increase in mantle viscosity. *Nature Geosci.* **8**, 311–314 (2015).

9. Immoor, J. et al. Evidence for {100}<011> slip in ferropericlase in Earth's lower mantle from high-pressure/high-temperature experiments. *Earth Planet. Sci. Lett.* **489**, 251–257 (2018).

10. Amodeo, J., Devincre, B., Carrez, P. H. & Cordier, P. Dislocation reactions, plastic anisotropy and forest strengthening in MgO at high temperature. *Mech. Materials* **71**, 62–73 (2014).

11. Reali, R. et al. Modeling plasticity of MgO by 2.5D dislocation dynamics simulations. *Mat. Science Eng. A* **690**, 52–61 (2017).

12. Stacey, F. & Davis, P. High pressure equations of state with applications to the lower mantle and core. *Phys. Earth Planet. Int.* **142**, 137–184 (2004).

13. Van Orman, J. & Crispin, K. Diffusion in oxides. *Rev. Mineral. Geochem.* **72**, 757–825 (2010).

14. Ita, J. & Cohen, R. E. Effects of pressure on diffusion and vacancy formation in MgO from non-empirical free-energy integrations. *Phys. Rev. Lett.* **79**, 3198 (1997).

15. Yang, M. & Flynn, C. Intrinsic diffusion properties of an oxide: MgO. *Phys. Rev. Lett.* **73**, 1809–1812 (1994).

16. Van Orman, J. A., Fei, Y., Hauri, E. H. & Wang, J. Diffusion in MgO at high pressures: constraints on deformation mechanisms and chemical transport at the core-mantle boundary. *Geophys. Res. Lett.* **30**, 1056–1059 (2003).

17. Landeiro Dos Reis, M., Giret, Y., Carrez, P. & Cordier, P. Efficiency of the vacancy pipe diffusion along an edge dislocation in MgO. *Comput. Mat. Sci.* **211**, 111490 (2022).

18. Ammann, M. W., Brodholt, J. P. & Dobson, D. P. Simulating diffusion. *Rev. Mineral. Geochem.* **71**, 201–224 (2010).

19. Karki, B. B. & Khanduja, G. Vacancy defects in MgO at high pressure. *Am. Mineral.* **91**, 511–516 (2006).

20. Runevall, O. & Sandberg, N. Self-diffusion in MgO—a density functional study. *J. Phys. Condens. Matter* **23**, 345402 (2011).

21. Reali, R. et al. The role of diffusion-driven pure climb creep on the rheology of bridgmanite under lower mantle conditions. *Sci Rep.* **9**, 2053 (2019).

22. Wang, Y. et al. High pressure, high-temperature deformation of $CaGeO_3$ (perovskite) ±MgO aggregates: implications for multiphase rheology of the lower mantle. *Geochem. Geophys. Geosyst.* **14**, 3389–3408 (2013).

23. Kaercher, P. et al. Two-phase deformation of lower mantle mineral analogs. *Earth Planet. Sci. Lett.* **456**, 134–145 (2016).

24. Kraych, A., Carrez P. H. & Cordier, P. On dislocation glide in $MgSiO_3$ bridgmanite at high-pressure and high-temperature. *Earth Planet. Sci. Lett.* **452**, 60–68 (2016).

25. Copley, S. & Pask, J. Plastic deformation of MgO single crystals up to 1600 C. *J. Am. Ceram. Soc.* **48**, 139–146 (1965).

26. Hulse, C. & Pask, J. Mechanical properties of magnesia single crystals in compression. *J. Am. Ceram. Soc.* **43**, 373–378 (1960).

27. Routbort, J. L. Work-hardening and creep of MgO. *Acta Metall.* **27**, 649–661 (1979).

28. Nzogang, B. et al. Characterization by scanning precession electron diffraction of an aggregate of bridgmanite and ferropericlase deformed at HP-HT. *Geochem. Geophys. Geosyst. B.* **19**, 582–594 (2018).

29. Ponte Castañeda, P. Second-order homogenization estimates for nonlinear composites incorporating field fluctuations: I—theory. *J. Mech. Phys. Solids* **50**, 737–757 (2002).

30. Madi, K., Forest, S., Cordier, P. & Boussuge, M. Numerical study of creep in two-phase aggregates with a large viscosity contrast: implications for the rheology of the lower mantle. *Earth Planet. Sci. Lett.* **237**, 223–238 (2005).

31. Kasemer, M., Zepeda-Alarcon, E., Carson, R., Dawson, P. & Wenk, H.-R. Deformation heterogeneity and intragrain lattice misorientation in high strength contrast, dual-phase bridgmanite/periclase. *Acta Mat.* **189**, 284–298 (2020).

32. Thielmann, M., Golabek, G. J. & Marquardt, H. Ferropericlase control of lower mantle rheology: impact of phase morphology. *Geochem. Geophys. Geosyst.* **21**, e2019GC008688 (2020).

33. Handy, M. R. Flow laws for rocks containing two nonlinear viscous phases: a phenomenological approach. *J. Struct. Geol.* **16**, 287–301 (1994).

34. Karki, B. B. et al. Structure and elasticity of MgO at high pressure. *Am. Mineral.* **82**, 51–60 (1997).

35. Hirth, J. P. & Lothe, J. *Theory of Dislocations* (John Wiley & Sons, 1992).

36. Singh, R. N. & Coble, R. L. Dynamic dislocation behavior in “pure” magnesium oxide single crystals. *J. Appl. Phys.* **45**, 981–989 (1974).

37. Boioli, F., Carrez, P., Cordier, P., Devincre, B. & Marquille, M. Modeling the creep properties of olivine by 2.5-dimensional dislocation dynamics simulations. *Phys. Rev. B* **92**, 014115 (2015).

38. Hershey, A. V. The elasticity of an isotropic aggregate of anisotropic cubic crystals. *J. Appl. Mech.* **21**, 236–240 (1954).

39. Kröner, E. Self-consistent scheme and graded disorder in polycrystal elasticity. *J. Phys. F: Metal Phys.* **8**, 2261–2267 (1978).

40. Yoo, H., Wuensch, B. J. & Petuskey, W. T. Oxygen self-diffusion in single crystal MgO: secondary-ion mass spectrometric analysis with comparison of results from gas–solid and solid–solid exchange. *Solid State Ion.* **150**, 207–221 (2002).

41. Ando, K., Kurokawa, Y. & Oishi, Y. Oxygen self-diffusion in Fe-doped MgO single crystals. *J. Chem. Phys.* **78**, 6890–6892 (1983).

42. Reali, R. et al. Modeling viscosity of (Mg,Fe)O at lowermost mantle conditions. *Phys. Earth Planet. Int.* **287**, 65–75 (2019).

43. Xu, S., Lin, J.-F. & Morgan, D. Iron partitioning between ferropericlase and bridgmanite in the Earth's lower mantle. *J. Geophys. Res. Solid Earth* **122**, 1074–1087 (2017).

44. Suquet, P. Overall properties of nonlinear composites: a modified secant moduli theory and its link with Ponte Castañeda nonlinear variational procedure. *C. R. Acad. Sci., Paris—Series IIB* **320**, 563–571 (1995).

45. Castelnau, O., Blackman, D. K., Lebensohn, R. A. & Ponte Castañeda, P. Micromechanical modeling of the viscoplastic behavior of olivine. *J. Geophys. Res.* **113**, B09202 (2008).

46. Raterron, P. et al. Multiscale modeling of upper mantle plasticity: from single-crystal rheology to multiphase aggregate deformation. *Phys. Earth Planet. Int.* **228**, 232–243 (2014).

47. Ponte Castañeda, P. Fully optimized second-order variational estimates for the macroscopic response and field statistics in viscoplastic crystalline composites. *Proc. Royal Soc. A.* **471**, 20150665 (2015).

48. Song, D. & Ponte Castañeda, P. Fully optimized second-order homogenization estimates for the macroscopic response and texture evolution of low-symmetry viscoplastic polycrystals. *Int. J. Plasticity* **110**, 272–293 (2018).

49. Castelnau, O. et al. Multiscale modeling of the effective viscoplastic behavior of Mg_2SiO_4 wadsleyite: bridging atomic and polycrystal scales. *Comptes Rendus Mécanique.* **348**, 827–846 (2020).

50. Budiansky, B. On the elastic moduli of some heterogeneous materials. *J. Mech. Phys. Solids* **13**, 223–227 (1965).

51. Gilormini, P., Brenner, R. & Castelnau, O. A similarity between the classical and modified secant extensions of the self-consistent model. *C. R. Acad. Sci., Paris—Series IIB* **329**, 523–527 (2001).

Acknowledgements

We acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 787198—TimeMan and from the National Science Foundation's (NSF) Collaborative Study of Earth's Deep Interior under EAR-1161046 and EAR-2009935. We are thankful to the CIDER programme set at the Kavli Institute for Theoretical Physics, University of California, Santa Barbara (NSF EAR−1135452, funded under the FESD Program).

Author information

Authors and Affiliations

1. Univ. Lille, CNRS, INRAE, Centrale Lille, UMR 8207 - UMET - Unité Matériaux et Transformations, Lille, France

   Patrick Cordier, Karine Gouriet, Timmo Weidner & Philippe Carrez

2. Institut Universitaire de France, Paris, France

   Patrick Cordier

3. Department of Earth, Environmental and Planetary Sciences, Case Western Reserve University, Cleveland, OH, USA

James Van Orman

4. Laboratoire PIMM, Arts et Metiers Institute of Technology, CNRS, CNAM, HESAM University, Paris, France

Olivier Castelnau

5. Seismological Laboratory, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA

Jennifer M. Jackson

Contributions

The study was designed by P. Cordier, K.G. and T.W. performed the dislocation dynamics calculations. O.C. performed the micromechanical (self-consistent) calculations. P. Cordier, K.G., T.W., J.V.O., O.C., J.M.J. and P. Carrez discussed and analysed the data. P. Cordier wrote the paper with contributions from all authors.

Corresponding author

Correspondence to Patrick Cordier.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Ian Jackson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Oxygen diffusion coefficients in periclase.

The black dotted line corresponds to the fit of Ita and Cohen[14] diffusion law at 0 GPa on the experimental data of Yoo et al.[40]. Comparison with results of Yang and Flynn[15] results is provided. The other lines (blue, green, red and magenta) correspond to this new set at 30, 60, 90 and 120 GPa respectively. The square symbols correspond to the diffusion coefficients along the geotherm deduced from this data. Comparison with data from Ita and Cohen[14] used in the main text (filled circles symbols) is provided

Source data

Extended Data Fig. 2 Pressure dependence of ionic diffusion.

The lines represent oxygen diffusion coefficients in periclase (Per). We compare the data of Ita and Cohen[14] (solid lines) to those obtained from the pressure dependence of Ita and Cohen[14] but anchored on data from Yoo et al.[40] (dotted lines). The symbols correspond to magnesium diffusion in bridgmanite (Bdm) considering a range of vacancy concentrations from 10^{-3} to 10^{-6}

Source data

Extended Data Fig. 3 Creep laws in bridgmanite (Bdm) and periclase (Per).

In this figure we illustrate the influence of possible uncertainties on the diffusion coefficients in bridgmanite and periclase. For bridgmanite (Bdm)

vacancy concentrations of 10^{-3} and 10^{-6} are considered. Creep in periclase is based on diffusion data of Ita and Cohen[14] and Yoo et al.[40]. Whatever the set of data considered, bridgmanite always deform faster than periclase

Source data

Extended Data Fig. 4 Viscosity contrast between bridgmanite (Bdm) and periclase (Per).

In this figure we illustrate the influence of possible uncertainties on the diffusion coefficients in bridgmanite and periclase. For bridgmanite (Bdm) vacancy concentrations in the range 10^{-3} to 10^{-6} are considered. Creep in periclase is based on diffusion data of Ita and Cohen[14] (Solid lines) and Yoo et al.[40] (dotted lines). Whatever the set of data considered, bridgmanite always deform faster than periclase

Source data

Extended Data Fig. 5 Pressure dependence of the aggregate rheology.

The flow stress \({\widetilde{\sigma }}_{0}\) (blue line) for a bridgmanite (Bdm) – periclase (Per) aggregate (volume fraction of MgO: 0.25), is obtained with the POSO Self-Consistent scheme, for vacancy concentrations of 10^{-3} and 10^{-6} in bridgmanite and two oxygen diffusion coefficients for periclase (from Ita and Cohen[14] and Yoo et al.[40]). Flow stresses of pure bridgmanite (red line) and pure periclase (black line) are also shown for comparison

Source data

Extended Data Fig. 6 Influence of periclase volume fraction (X_v = 10^{-3}).

Behavior for a vacancy concentration in bridgmanite of $X_v = 10^{-3}$ and oxygen diffusion coefficient in periclase from Ita & Cohen[14] at various pressures. Left: Effective flow stress \({\widetilde{\sigma }}_{0}\) based on four homogenization procedures: Reuss, Voigt, Hashin-Strikman (HS+) and the POSO extension of the Self-Consistent scheme (SC). Right: Relative strain-rates in bridgmanite and periclase.

Source data

Extended Data Fig. 7 Influence of periclase volume fraction ($X_v = 10^{-6}$).

Behavior for a vacancy concentration in bridgmanite of $X_v = 10^{-6}$ and oxygen diffusion coefficient in periclase from Ita & Cohen[14] at various pressures. Left: Effective flow stress \({\widetilde{\sigma }}_{0}\) based on four homogenization procedures: Reuss, Voigt, Hashin-Strikman (HS+) and the POSO extension of the Self-Consistent scheme (SC). Right: Relative strain-rates in bridgmanite and periclase.

Source data

Extended Data Table 1 Additional diffusion coefficients
Extended Data Table 2 Periclase parameters

Source data

Source Data Fig. 1

Source Data Fig. 2

Source Data Fig. 3

Source Data Fig. 4

Source Data Extended Data Fig. 1

Source Data Extended Data Fig. 2

Source Data Extended Data Fig. 3

Source Data Extended Data Fig. 4

Source Data Extended Data Fig. 5

Source Data Extended Data Fig. 6

Source Data Extended Data Fig. 7

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Cordier, P., Gouriet, K., Weidner, T. *et al.* Periclase deforms more slowly than bridgmanite under mantle conditions. *Nature* **613**, 303–307 (2023).

https://doi.org/10.1038/s41586-022-05410-9

- Received: 10 February 2022
- Accepted: 03 October 2022
- Published: 11 January 2023
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05410-9

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

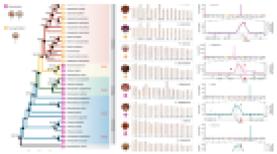
Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

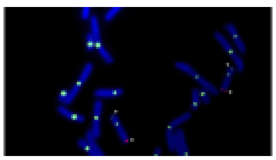
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05410-9



Ancient homomorphy of molluscan sex chromosomes sustained by reversible sex-biased genes and sex determiner translocation

24 October 2022

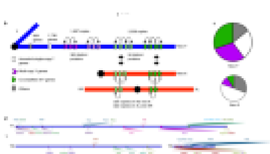
Wentao Han, Liangjie Liu, … Shi Wang



Miniscule differences between sex chromosomes in the giant genome of a salamander

14 December 2018

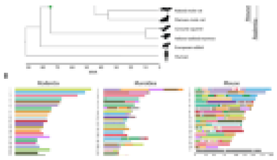
Melissa C. Keinath, Nataliya Timoshevskaya, … Jeramiah J. Smith



Massive gene amplification on a recently formed Drosophila Y chromosome

28 October 2019

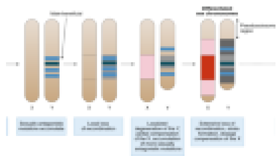
Doris Bachtrog, Shivani Mahajan & Ryan Bracewell



[3D chromatin remodelling in the germ line modulates genome evolutionary plasticity]

11 May 2022

Lucía Álvarez-González, Frances Burden, … Aurora Ruiz-Herrera



[Molecular and evolutionary dynamics of animal sex-chromosome turnover]

25 November 2019

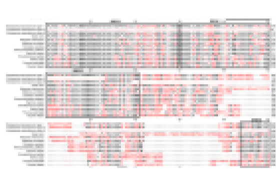
Beatriz Vicoso



[Rapid evolutionary dynamics of an expanding family of meiotic drive factors and their hpRNA suppressors]

03 December 2021

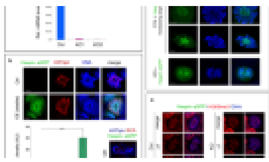
Jeffrey Vedanayagam, Ching-Jung Lin & Eric C. Lai



The germ cell marker dead end reveals alternatively spliced transcripts with dissimilar expression

20 February 2019

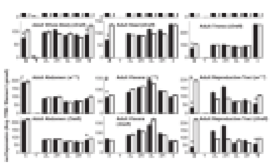
Ana Carina Nogueira Vasconcelos, Danilo Pedro Streit Jr., … Goro Yoshizaki



Distinct roles of haspin in stem cell division and male gametogenesis

06 October 2021

Katerina Soupsana, Eleftheria Karanika, … Spyros Georgatos



Dynamic sex chromosome expression in Drosophila male germ cells

09 February 2021

Sharvani Mahadevaraju, Justin M. Fear, … Brian Oliver

- Article
- Open Access
- Published: 21 December 2022

The molecular evolution of spermatogenesis across mammals

- Florent Murat ,
- Noe Mbengue ,
- Sofia Boeg Winge,
- Timo Trefzer,
- Evgeny Leushkin,
- Mari Sepp,
- Margarida Cardoso-Moreira,
- Julia Schmidt,
- Celine Schneider,
- Katharina Mößinger,
- Thoomke Brüning,
- Francesco Lamanna,
- Meritxell Riera Belles,
- Christian Conrad,
- Ivanela Kondova,
- Ronald Bontrop,
- Rüdiger Behr,
- Philipp Khaitovich,
- Svante Pääbo,
- Tomas Marques-Bonet,
- Frank Grützner,
- Kristian Almstrup,
- Mikkel Heide Schierup &
- …
- Henrik Kaessmann

Show authors

Nature **volume 613**, pages 308–316 (2023)

- 15k Accesses

- 156 Altmetric

- Metrics details

Abstract

The testis produces gametes through spermatogenesis and evolves rapidly at both the morphological and molecular level in mammals[1,2,3,4,5,6], probably owing to the evolutionary pressure on males to be reproductively successful[7]. However, the molecular evolution of individual spermatogenic cell types across mammals remains largely uncharacterized. Here we report evolutionary analyses of single-nucleus transcriptome data for testes from 11 species that cover the three main mammalian lineages (eutherians, marsupials and monotremes) and birds (the evolutionary outgroup), and include seven primates. We find that the rapid evolution of the testis was driven by accelerated fixation rates of gene expression changes, amino acid substitutions and new genes in late spermatogenic stages, probably facilitated by reduced pleiotropic constraints, haploid selection and transcriptionally permissive chromatin. We identify temporal expression changes of individual genes across species and conserved expression programs controlling ancestral spermatogenic processes. Genes predominantly expressed in spermatogonia (germ cells fuelling spermatogenesis) and Sertoli (somatic support) cells accumulated on X chromosomes during evolution, presumably owing to male-beneficial selective forces. Further work identified transcriptomal differences between X- and Y-bearing spermatids and uncovered that meiotic sex-chromosome inactivation (MSCI) also occurs in monotremes and hence is common to mammalian sex-chromosome systems. Thus, the mechanism of meiotic silencing of unsynapsed chromatin, which underlies MSCI, is an ancestral mammalian feature. Our study illuminates the molecular evolution of spermatogenesis and associated selective forces, and provides a resource for investigating the biology of the testis across mammals.

Main

The rapid evolution of the testis across mammals is probably mainly explained by positive selection associated with sperm competition, which reflects the evolutionary pressure on males to achieve reproductive success[7].

Consequently, testis sizes, sperm production rates, sperm morphologies and other cellular traits substantially vary across mammals, even between closely related species such as the great apes, due to great mating system differences, especially regarding the extent of female promiscuity[7]. The rapid evolution of the testis is reflected at the molecular level. Previous gene expression comparisons for various organs across mammals revealed that rates of evolutionary expression change are highest in the testis, probably due to frequent adaptive changes but potentially also widespread relaxation of purifying selection[1,2,3,4,5,6]. Consistently, genes with testis-specific expression tend to be enriched with genes whose coding sequences have been shaped by positive selection[8]. In addition, new genes that emerge during evolution tend to be predominantly expressed in the testis and thus probably also contribute to its rapid phenotypic evolution[3,9].

The testis also shows several other unique molecular features. First, chromatin in spermatogenic cells is massively remodelled during spermatogenesis, a process that culminates in the tight packaging of DNA around protamines in the compact sperm head[10]. This remodelling leads to widespread leaky transcription in the genome[11], which in turn probably facilitates the initial transcription and, hence, the frequent emergence of new testis-expressed genes and alternative exons during evolution[3,9,11,12]. Second, the differentiation of sex chromosomes from ancestral autosomes triggered the emergence of MSCI in eutherians and marsupials[13] (therians), which led to the establishment of backup gene copies that substitute for parental genes on the X during meiosis[3,14]. In spite of MSCI, the X chromosome has become enriched with testis-expressed genes during evolution[3,15,16,17,18,19,20], presumably due to sexually antagonistic selective forces favouring the fixation of male-beneficial mutations on this chromosome[21]. Finally, translational regulation of transcriptomes is widespread across spermatogenesis[6].

Previous large-scale transcriptomic investigations of testis evolution were largely limited to bulk-organ samples[1,2,3,4,5,6,19]. Recent high-throughput single-cell (sc) or single-nucleus (sn) RNA-sequencing (RNA-seq) technologies enable detailed investigations of the cellular and molecular evolution of the testis, as exemplified by two scRNA-seq comparisons

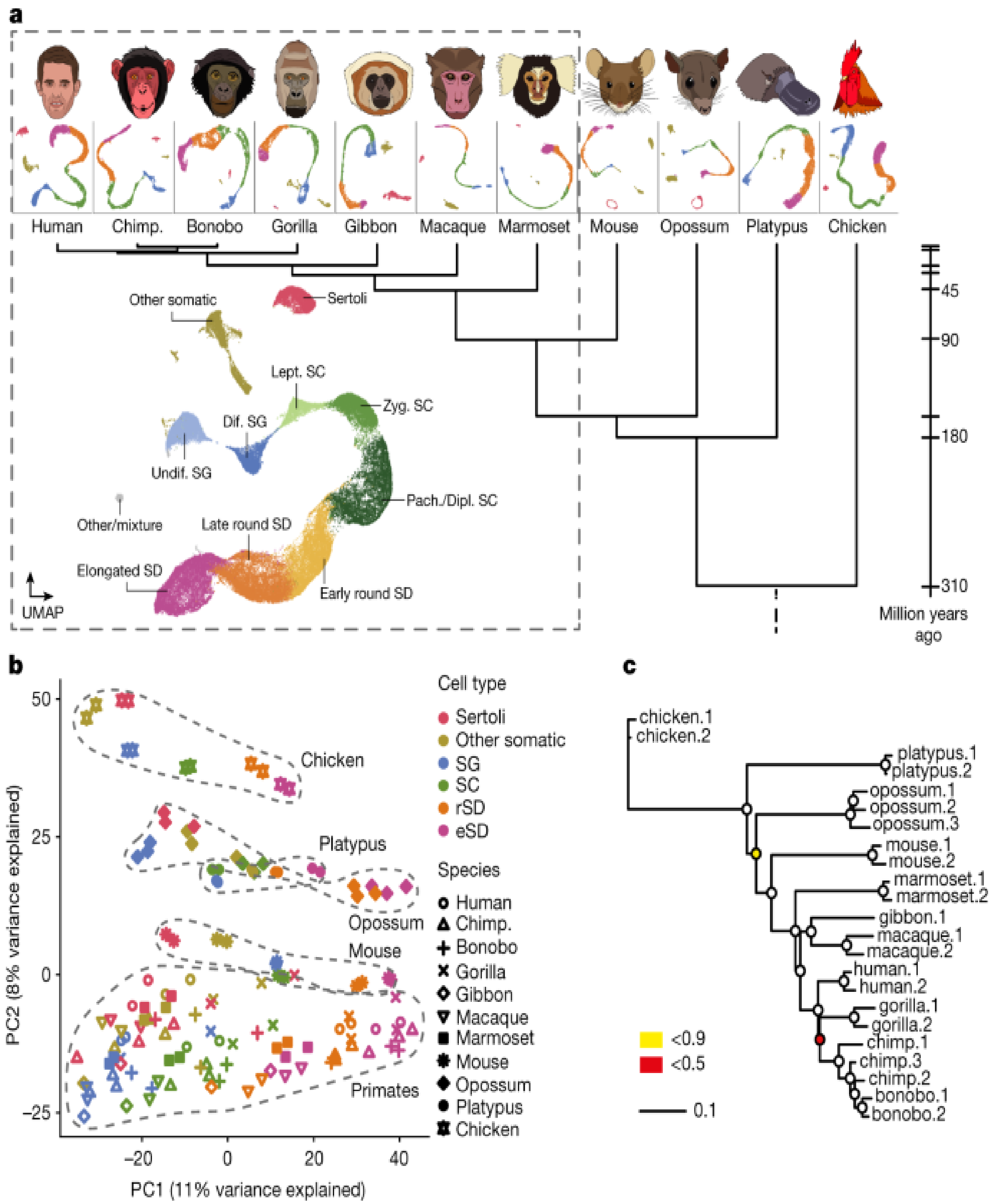
between human, macaque and mouse[22,23], but a comprehensive investigation of the evolution of spermatogenesis across all main mammalian lineages is lacking.

Here we provide an extensive snRNA-seq resource covering testes from ten representative mammals and a bird, enabling detailed comparisons of spermatogenic cells and underlying gene expression programs within and across mammals (https://apps.kaessmannlab.org/SpermEvol/). Our evolutionary analyses of these data unveiled ancestral as well as species- and lineage-specific cellular and molecular characteristics of mammalian spermatogenesis.

Spermatogenesis across 11 species

We generated snRNA-seq data for testes from ten species that cover the three main mammalian lineages and include key primate species, representing all simian (anthropoid) lineages (Fig. 1a): eutherian mammals (representatives for five of the six extant ape lineages, including humans; rhesus macaque, an Old World monkey; common marmoset, a New World monkey; and mouse), marsupials (grey short-tailed opossum) and egg-laying monotremes (platypus). Corresponding data were generated for a bird (red jungle fowl, the progenitor of domestic chicken; hereafter referred to as ‘chicken’), to be used as an evolutionary outgroup. The dataset consists of 27 libraries, with one to three biological replicates per species and a median of roughly 275 million snRNA-seq reads per library (Supplementary Table 1). We refined and extended existing genome annotations across all species on the basis of bulk-testis RNA-seq data (seven libraries) (Supplementary Tables 1 and 2 and Methods), to ensure optimal read-mapping and prevent biases in cross-species analyses. After quality controls and filtering steps (Methods), we obtained transcriptomes for a total of 97,521 high-quality nuclei for the 11 species, with a mean of 8,866 cells per species, a median of 1,856 RNA molecules (unique molecular identifiers (UMIs)) detected per cell and low percentages of mitochondrial UMIs (Supplementary Fig. 1 and Supplementary Table 1).

Fig. 1: snRNA profiling across ten mammals and a bird.



a, Species sampled and uniform manifold approximation and projection (UMAP) of snRNA-seq datasets. UMAP of the integrated primate dataset (dashed box), showing undifferentiated and differentiated SG (undif. SG and dif. SG, respectively), leptotene, zygotene, pachytene and diplotene SCs (lept. SC, zyg. SC, pach. SC and dipl. SC, respectively), spermatids (SD)

and somatic cell types. Chimp., chimpanzee. **b**, Principal component (PC) analysis of cell-type pseudo-bulks. Species and lineages are encircled by a dashed line. Each symbol represents an individual. **c**, Gene expression phylogeny based on pseudo-bulk transcriptomes for whole testes. Bootstrap values (4,498 1:1 orthologous amniote genes were randomly sampled with replacement 1,000 times) are indicated by circles, ≥0.9 (white fill).

We identified the main germ cell types[11] along the continuous cellular proliferation and differentiation path of spermatogenesis across all species: spermatogonia (SG), the mitotic cells fuelling spermatogenesis, including spermatogonial stem cells; spermatocytes (SC), where meiosis takes place; and the haploid round spermatids (rSD) and elongated spermatids (eSD), which together reflect spermiogenesis (Fig. 1a, Extended Data Figs. 1a–f, Supplementary Table 3 and Methods). We also identified separate clusters corresponding to somatic testicular cells, in particular Sertoli cells (except for platypus, in which Sertoli cells could not be unambiguously distinguished), the main spermatogenesis support cells, but also other cell types, such as Leydig cells, peritubular cells, endothelial cells and macrophages (Extended Data Figs. 1a–e). The close evolutionary relationship of the seven primates in our study enabled the direct integration of datasets across these species and thus the identification of sub-cell types that correspond to intermediate events during spermatogenesis (Fig. 1a). For all species, we traced the dynamic gene expression programs underlying spermatogenic differentiation and key molecular events, thus also identifying a host of new marker genes (Extended Data Fig. 1g and Supplementary Table 4).

To obtain an overview of cell-type relationships across species, we performed a principal component analysis (PCA) based on pseudo-bulk cell-type transcriptomes (Fig. 1b). The first principal component (PC1) orders the spermatogenic cell types according to the progression of spermatogenesis for all species (Fig. 1b, from left to right). This observation suggests that our data capture ancestral aspects of spermatogenic gene expression programs that are shared across mammals or amniotes, despite the rapid evolution of the testis[1,2,3,4,5,6] and the long divergence times of 310 million years (Fig. 1a). PC2 separates the data by species or lineages, reflecting diverged aspects of spermatogenesis, whereas PC1 and PC2 together separate somatic and spermatogenic cell types for each species. The

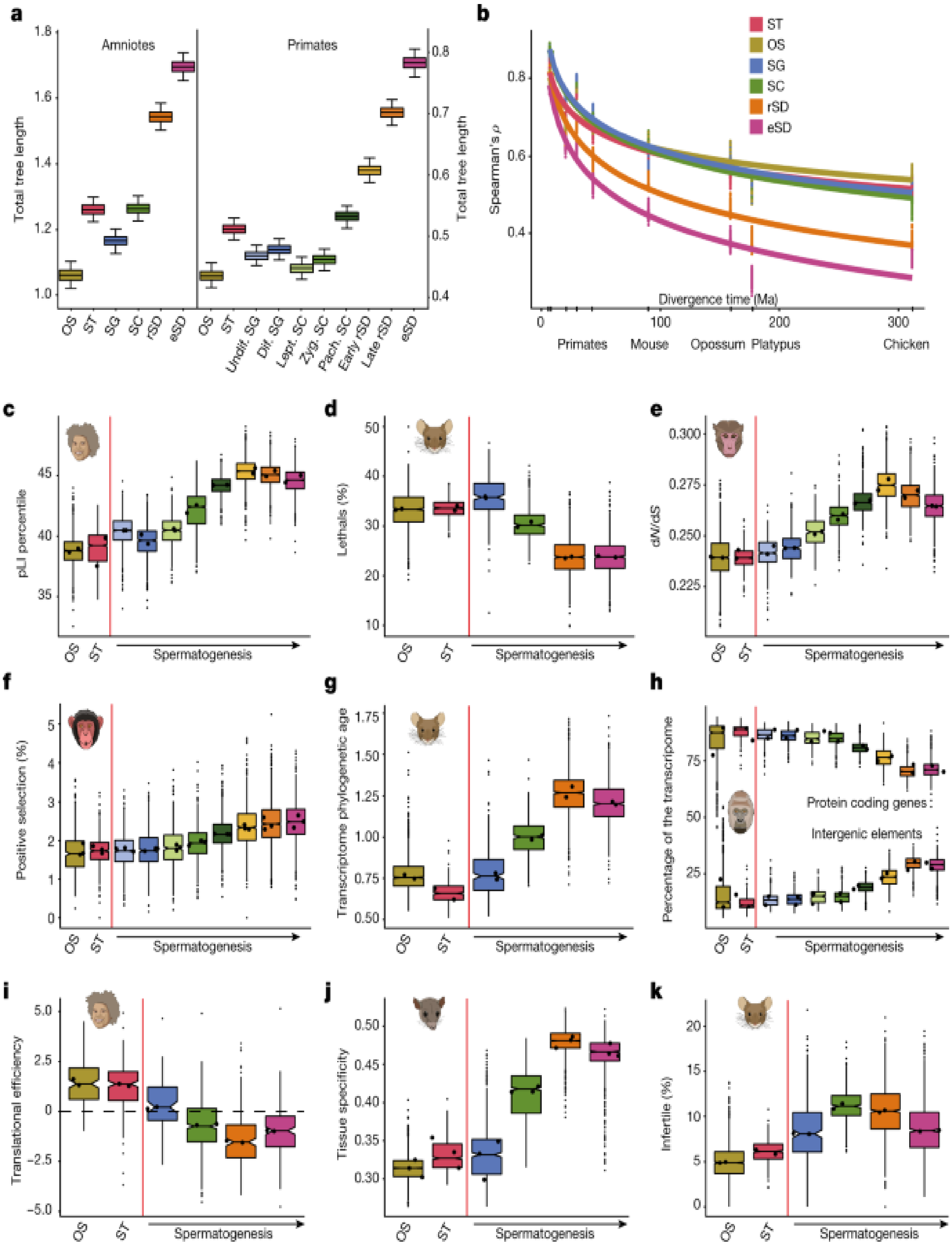
close clustering of biological replicates is a further indicator of the high data quality.

Rates of evolution along spermatogenesis

To investigate rates of gene expression evolution across cell types, we reconstructed gene expression trees (Methods). A tree based on pseudo-bulk transcriptomes for the whole testis (Fig. 1c) recapitulates the known mammalian phylogeny (Fig. 1a, except for the gibbon–macaque grouping), akin to previous trees based on bulk-tissue RNA-seq data across mammalian organs[1,2]. This observation is consistent with the view that regulatory changes steadily accumulated over evolutionary time[1], with present-day RNA abundances reflecting the evolution of mammalian lineages and species.

To trace the cellular source of the rapid evolution of the testis, we built expression trees for the different cell types, which also recapitulate the known species relationships (Extended Data Fig. 2a). Notably, the total branch lengths of the trees, which reflect the amount of evolutionary expression change, vary substantially between cell types (Fig. 2a). Whereas the rate of expression evolution is similar in Sertoli cells and diploid spermatogenic cells (and lower than that in other somatic cell types), it is substantially higher in the postmeiotic haploid cell types (rSD and eSD), consistent with a recent inference based on data for three eutherians[23]. The higher resolution afforded by a primate-specific analysis provides further details (Fig. 2a). Starting in late meiosis (pachytene SC), evolutionary rates progressively increase until the end of spermiogenesis (late eSD). Thus, late spermatogenic stages drive the previously observed rapid evolution of the testis[1,2,3,4,5,6].

Fig. 2: Gene expression divergence and evolutionary forces.



a, Total branch lengths of expression trees among testicular cell types for amniotes and primates. Box plots show the median (central value); upper and lower quartile (box limits) and 95% confidence intervals (whiskers) for

1,000 bootstrap replicates. **b**, Spearman's correlations between humans and other species from 100 bootstrap replicates (dots). Lines correspond to linear regression trends (after log transformation of the time axis. Regression R^2 values range from 0.86 to 0.97. Ma, millions of years ago. **c**, Mean pLI value of expressed genes in human (the pLI score reflects the tolerance of a gene to a loss-of-function mutation; lower values mean less tolerance). **d**, Percentage of expressed genes leading to a lethal phenotype when knocked out in mouse (out of 4,742 knockouts[25]). **e**, Mean normalized ratio of nonsynonymous (d*N*) over synonymous (d*S*) nucleotide substitutions of expressed genes in macaque. **f**, Percentage of expressed genes under positive selection (out of 11,170 genes tested for positive selection) in chimpanzee. **g**, Mean phylogenetic age of expressed genes in mouse. **h**, Percentage of UMIs mapping to protein-coding genes (top) or intergenic elements (bottom) in gorilla. **i**, Translational efficiency values (data from ref. [6]) are plotted for all genes with predominant expression in a given cell type in human. **j**, Mean of tissue-specificity values (data from ref. [2]) in opossum. **k**, Percentage of expressed genes associated with infertility (out of 3,552 knockouts) in mouse. **c**–**h**,**j**,**k**, Plotted is the mean value per cell. **c**–**k**, Superimposed thick black dots indicate medians from biological replicates. Box plots depict the median (centre value); upper and lower quartile (box limits) with whiskers at 1.5 times the interquartile range. Red lines separate somatic (OS, other somatic; ST, Sertoli cells) and germ cells. Data for other studied species are shown in Extended Data Fig. 3.

Pairwise species comparisons, including downsampling analyses (Extended Data Fig. 2b), confirm the rapid expression evolution of postmeiotic cell types across amniotes and that gene expression divergence increases with evolutionary time (Fig. 2b), in accord with the expression phylogeny results (Extended Data Fig. 2a). It is noteworthy, however, that expression divergence levels are roughly as similar between human and chicken as they are between human and platypus, although the bird lineage diverged 110 million years before the separation of monotremes and therians (that is, eutherians and marsupials). This observation, previously made for whole organs[1], supports at the cellular level the notion that the conservation of core spermatogenic functions restricts transcriptome divergence.

Evolutionary forces

We sought to trace the evolutionary forces underlying the rapid evolution of late spermatogenesis. Two non-mutually exclusive patterns of natural selection may account for this observation. First, later stages of spermatogenesis might evolve under weaker purifying selection (that is, reduced functional constraints) and hence be less refractory to change. Second, the greater divergence in later stages might result from stronger positive selection, increasing the rate of fixation of adaptive changes. To investigate patterns of functional constraint during spermatogenesis, we assessed the tolerance to functional mutations of genes[24] used in different spermatogenic stages in humans, which showed a progressive increase of mutational tolerance starting during meiosis and culminating in early spermiogenesis (Fig. 2c and Extended Data Fig. 3a). Consistently, on the basis of a set of neutrally ascertained mouse knockouts[25], we found that the percentage of expressed genes associated with lethality decreases during spermatogenesis (Fig. 2d). Also, in agreement with a progressive reduction in functional constraints towards later spermatogenic stages, we find that the normalized rate of amino acid altering substitutions in coding sequences across primates is higher in late spermatogenesis (Fig. 2e and Extended Data Fig. 3b), although this increase might additionally reflect a higher proportion of genes under positive selection. Indeed, an examination of the temporal expression pattern of genes whose encoded protein sequences have been shaped by positive selection revealed a notable increase in percentages of positively selected genes used during spermatogenesis, with a peak in rSD (Fig. 2f and Extended Data Fig. 3c).

Because new genes also contribute to evolutionary innovations, we investigated the temporal contribution of recently emerged genes to gene expression programs in germ cells, using an index that combines the phylogenetic age of genes with their expression[2] (Methods). This analysis revealed that transcriptomes become younger during spermatogenesis (Fig. 2g and Extended Data Fig. 3d), indicating that new genes have increasingly more prominent roles in later stages, particularly in rSD, consistent with previous observations[22,23]. Previous work based on bulk cell-type analyses in mouse[11] uncovered a transcriptionally permissive chromatin environment during spermatogenesis, in particular in rSD, which was suggested to have facilitated the emergence of new genes during evolution[3,9,11]. Consistently, we detect in all species considerably increased contributions of intergenic

transcripts after meiosis and a concomitant decrease in the contributions of protein-coding genes (Fig. 2h and Extended Data Fig. 3e). Notably, an analysis of translatome data[6] revealed a decline of translational efficiencies of transcripts during spermatogenesis (reaching a minimum in rSD) in all species (Fig. 2i and Extended Data Fig. 3f). This decline is consistent with observations from mouse bulk data for a restricted number of cell types[6] and probably mitigates the functional consequences of the concurrent increase in transcriptional promiscuity of the genome.

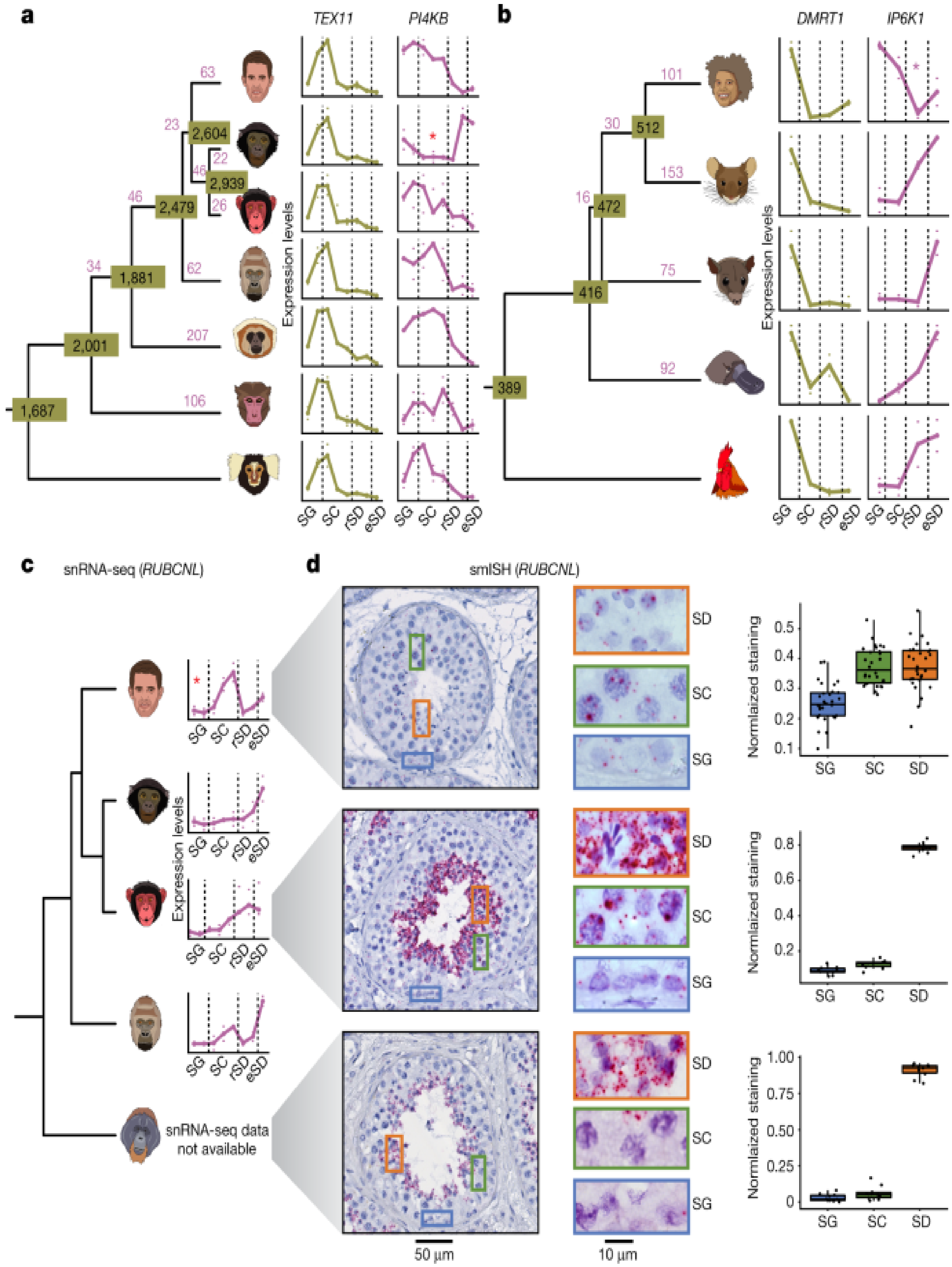
We next explored the reasons underlying the dynamic changes of selective forces and patterns of innovation during spermatogenesis. The breadth of expression across tissues and developmental processes (here referred to as expression pleiotropy) was proposed to represent a key determinant of the types of mutation that are permissible under selection[26]. We, therefore, assessed patterns of expression pleiotropy across spermatogenesis using spatiotemporal transcriptome data for several mammalian organs[2], which revealed that genes used later in spermatogenesis, in particular those in rSD, have substantially more specific spatiotemporal profiles than genes used earlier in spermatogenesis and in somatic cells (Fig. 2j and Extended Data Fig. 3g,h). Given that a decrease in expression pleiotropy can explain both a decrease in functional constraints and an increase in adaptation[2,26], we suggest that it is probably a main contributor to the accelerated molecular evolution in late spermatogenesis. In addition, the specific type of selection acting on haploid cells[27] (haploid selection), in which expressed alleles are directly exposed to selection, may have contributed to the exceptionally rapid evolution of rSD.

Whereas the tissue- and time-specific late spermatogenic genes, in general, are not essential for viability (Fig. 2d,j and Extended Data Fig. 3g,h, above), we proposed that the specific aforementioned evolutionary forces indicate that many of these genes evolved crucial roles in spermatogenesis. Indeed, we find that the proportion of genes associated with infertility[25] is relatively high in SC and spermatids (especially rSD): higher than in SG and somatic cells (Fig. 2k).

Gene expression conservation and innovation

We next sought to trace the individual genes underlying conserved (ancestral) and diverged aspects of germ cells by comparing expression trajectories along spermatogenesis of one-to-one (1:1) orthologous genes across species[2] (Supplementary Fig. 2 and Methods). For the primates, this analysis revealed roughly 1,700–2,900 genes with conserved expression trajectories across different lineages or species (Fig. 3a and Supplementary Table 5). For example, the temporal expression of 1,687 genes is conserved across the seven primates and probably reflects the core ancestral gene expression program of the simian testis.

Fig. 3: Evolution of gene expression trajectories along spermatogenesis.



a,**b**, The numbers of changed trajectories (in purple) and conserved trajectories (in olive) are indicated. In **a**–**c**, red asterisks indicate the branch

for which a trajectory change has been called. For each replicate, mean expression levels across cells of a given cell type were calculated. Marks indicate values for the replicates with the highest and lowest mean expression levels, dots indicate the median of mean expression values of the replicates (three replicates for chimpanzee and opossum; two for human, bonobo, gorilla, macaque, marmoset, mouse, platypus and chicken; one for gibbon). In **a**–**c**, dashed vertical lines separate SG, SCs, rSD and eSD. **a**, Primate trajectories (human, bonobo, chimpanzee, gorilla, gibbon, macaque and marmoset; from top to bottom) based on 4,459 1:1 orthologues. *TEX11* and *PI4KB* are examples of conserved and changed trajectories, respectively. **b**, Amniote trajectories (human, mouse, opossum, platypus and chicken; from top to bottom) based on 2,927 1:1 orthologues. *DMRT1* and *IP6K1* are examples of conserved and changed trajectories, respectively. **c**, *RUBCNL* expression trajectory from snRNA-seq data along spermatogenesis for human, bonobo, chimpanzee and gorilla (from top to bottom). Available orangutan samples did not allow for the generation of high-quality snRNA-seq data. **d**, Detection of *RUBCNL* (each red dot reflects a single transcript) expression in human, chimpanzee and orangutan testis by smISH using RNAScope. Left, seminiferous tubule cross sections counterstained with haematoxylin, and closeups on areas containing SG, SC and spermatids (SD). Right, quantification of *RUBCNL* expression levels in ten tubules per section ($n = 3$ for human, $n = 1$ for chimpanzee and orangutan). Box plots represent the mean (central value) distribution of staining (dots); upper and lower quartile (box limits) with whiskers at 1.5 times the interquartile range.

By contrast, we detected 635 trajectory changes during the evolution of apes and Old World monkeys (Fig. 3a and Supplementary Table 5). For example, 63 and 94 trajectory changes occurred in the human and chimpanzee–bonobo lineages, respectively, since their divergence roughly 7 million years ago. We spatially validated three of these changes using single-molecule RNA in situ hybridization (smISH) for three great apes (human, chimpanzee and orangutan as the outgroup) (Fig. 3c, Extended Data Fig. 4, Supplementary Fig. 3, Supplementary Table 6 and Methods). Our smISH experiments confirm that the expression of the gene *RUBCNL*, encoding a regulator of autophagy[28], changed in the human lineage, with a relative reduction of expression levels in spermatids (Fig. 3c,d). Thus, the role of *RUBCNL* in autophagy during spermatogenesis[29] is potentially different in

humans compared to other Old World anthropoids. We also confirmed the inversion of the expression trajectory of the myosin-encoding gene *MYO3B* in SG and rSD on the chimpanzee–bonobo lineage (Extended Data Fig. 4a), as well as the expression increase of *ADAMTS17*, a family-member of proteases with key spermatogenic functions[30], in rSD relative to SG in humans (Extended Data Fig. 4b). We note that observed quantitative and partly qualitative expression pattern differences between the two complementary data types are expected because of technical differences.

We also assessed the conservation of expression trajectories across the three main mammalian lineages and amniotes (Fig. 3b and Supplementary Table 7). The 416 genes with conserved expression across all mammals probably trace back to ancestral gene expression programs of the ancestral mammalian testis, whereas 389 genes may represent the core ancestral spermatogenic program of amniotes. A notable example is *DMRT1*, which is highly expressed specifically in SG across amniotes and is required for mouse spermatogonial stem cell maintenance and replenishment[31].

In agreement with these highly conserved sets of genes having key roles in spermatogenesis, our analyses of fertility phenotypes[25] unveiled that genes involved in fertility are significantly more conserved in their expression trajectories than genes not associated with fertility (Extended Data Fig. 5b). Consistently, a Gene Ontology[32] enrichment analysis indicates an involvement of conserved genes in fundamental spermatogenic processes that are typical of the cell type in which they show peak expression (Extended Data Fig. 6a,b). Thus, genes with conserved trajectories for which spermatogenesis functions remain uncharacterized represent promising candidates for the exploration of fertility phenotypes (Supplementary Tables 5, 7 and 8). Notably, genes with lineage-specific trajectory changes are enriched with broader, typically metabolic, processes (Extended Data Fig. 6c,d). They are also significantly more tissue- and time-specific than genes with conserved expression (Extended Data Fig. 6e), which may have facilitated their expression change during evolution because of reduced pleiotropic constraints. However, genes with changed trajectories nevertheless include many genes for which key fertility functions have been described (Supplementary Tables 7 and 8). For example, *IP6K1*, which elicits infertility when knocked out in the mouse[33], shows strongly

increasing expression towards the end of spermiogenesis in all amniotes except the primates, for which expression is high in SG and then declines (Fig. 3b and Extended Data Fig. 5c). Thus, the primary function of *IP6K1* probably shifted from late to early spermatogenesis during primate evolution.

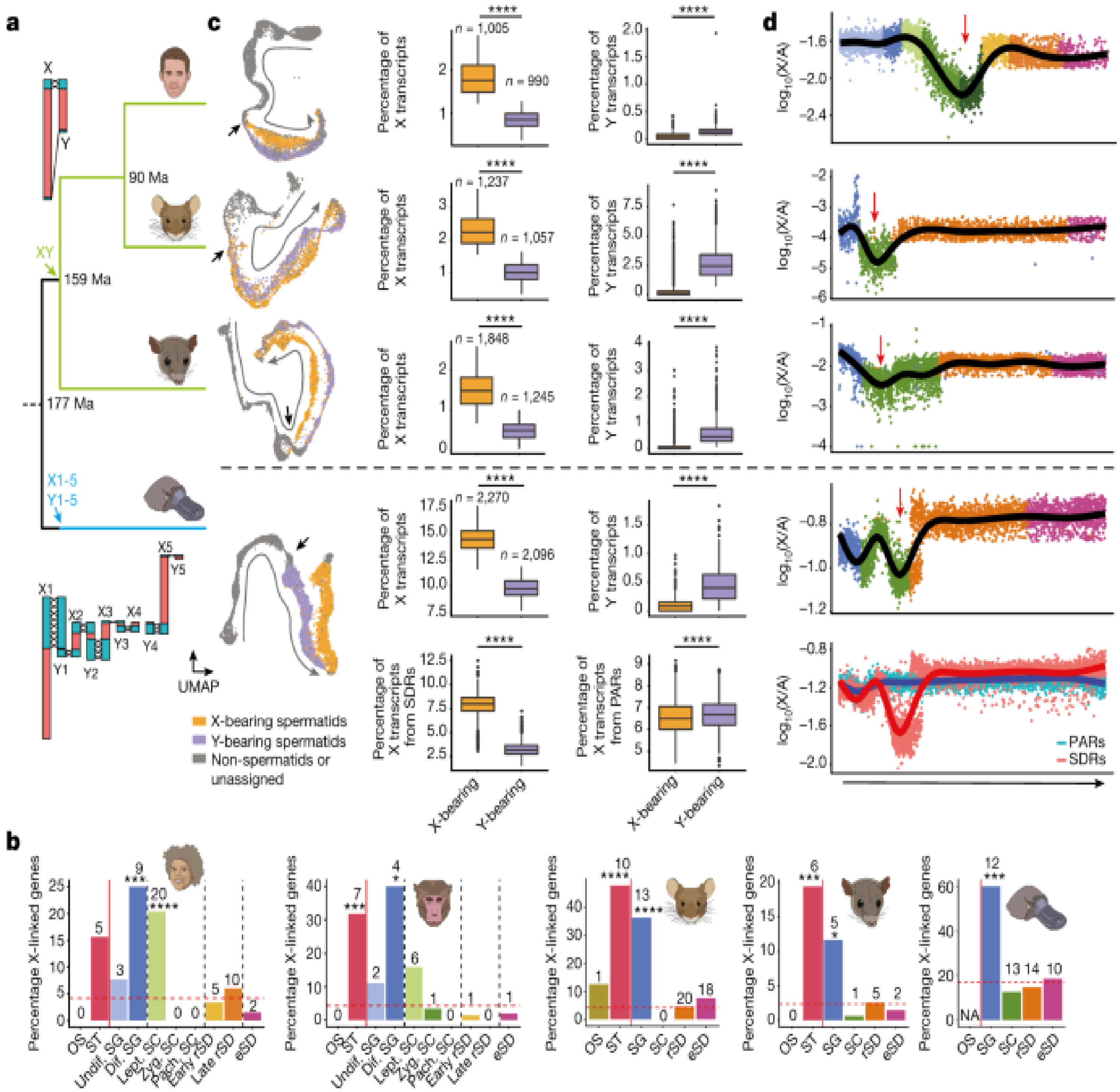
In conjunction with mouse fertility data, we used our data to explore the contribution of new genes (mostly arising from gene duplications) to the evolution of new spermatogenic functions, focusing on the rodent lineage leading to mouse. This analysis revealed the emergence of key spermatogenic genes at different time points during evolution (Extended Data Fig. 7a and Supplementary Table 8), such as two rodent-specific retrogenes—*D1Pas1* and *H2al2a*—that originated through RNA-based duplication from parental genes on the X chromosome (Extended Data Fig. 7c). *D1Pas1* is essential for meiosis[34], whereas *H2al2a* is essential for genome compaction in late spermatogenesis[35]. Both genes show strongly increasing expression levels in late spermatogenesis (Extended Data Fig. 7b) and are thus in agreement with new genes contributing to functional roles predominantly during late spermatogenesis (above, Fig. 2g and Extended Data Fig. 3d).

Finally, we used our data to investigate ligand–receptor interactions underlying the communication between Sertoli cells, which have a central role in supporting and controlling spermatogenesis[11], and germ cells across species (Methods). Our cross-species comparisons revealed various conserved known and new ligand–receptor interactions (Extended Data Fig. 8a,b and Supplementary Table 9). For example, our data provide evidence that communication of Sertoli cells with SG, SC and rSD occurs in all amniotes through interactions of the cell adhesion molecule CADM1 (ref. [36]) or between CADM1 (Sertoli cells) and NECTIN3 (SC and rSD) (Supplementary Table 9). We note that CADM1 was previously thought to not to be expressed in Sertoli cells[37], but our data show high expression in this cell type across amniotes, in agreement with human protein atlas data (proteinatlas.org/ENSG00000182985-CADM1/tissue/testis#). Our work also supports the notion that the NECTIN2–NECTIN3 complex mediates the communication of Sertoli cells with spermatids not only in mouse[38] but also in humans[39] (Supplementary Table 9).

Sex chromosomes

Sex chromosomes emerged twice in parallel in mammals from different sets of ancestral autosomes. The therian XY chromosome system originated just before the split of eutherians and marsupials (Fig. 4a) and hence has evolved largely independently in these two lineages. Around the same time, the monotreme sex-chromosome system arose from a different pair of autosomes and subsequently expanded to five XY pairs[40]. These sex-chromosome formation events entailed substantial remodelling of gene contents and expression patterns due to structural changes and sex-related selective forces[3]. We used our data to systematically assess testicular expression patterns of sex chromosomal genes and their evolution across mammals.

Fig. 4: Mammalian sex-chromosome evolution.



a, Phylogenetic tree for human, mouse, opossum and platypus (from top to bottom). Arrows show sex systems origination[40]. Illustration of therian (human, top) and monotreme (platypus, bottom) sex chromosomes. Recombining (crosses) PARs in turquoise and SDRs in red. **b**, Percentages of X-linked genes among testis-specific genes with predominant expression in a given cell type for human, macaque, mouse, opossum and platypus (from left to right; Supplementary Table 10). The red dashed line represents the expected percentages of X-linked genes, if testis-specific genes with predominant expression in the different cell types were randomly distributed across the genome. Asterisks indicate significance after two-sided exact binomial testing (Benjamini–Hochberg corrected *P* values from left to right:

0.000353, 1.18×10^{-7}, 0.000687, 0.0142, 1.08×10^{-7}, 3.45×10^{-8}, 0.000792, 0.0355 and 0.00043). The number of testis-specific X-linked genes enriched in each cell type is indicated above each bar. **c**, UMAP representation of germ cells (left). Progression of spermatogenesis (grey arrow) and meiotic divisions (black arrows) are indicated. Spermatids identified as X- and Y-bearing are coloured in orange and purple, respectively. Box plots show the median (central value) percentages of X and Y transcripts in X- and Y-bearing spermatids; upper and lower quartile (box limits) with whiskers at 1.5 times the interquartile range. Two-sided Wilcoxon rank-sum tests were performed for statistical comparisons (Benjamini–Hochberg corrected ****$P < 2.2 \times 10^{-16}$). **d**, X-to-autosome transcript ratios in individual germ cells across spermatogenesis (from left to right). For spermatids, only X-bearing cells are considered. Lines depict generative additive model trend. Red arrows indicate MSCI. **c**,**d**, For platypus, X transcripts are dissected according to their location on the X chromosomes (Methods).

We first contrasted cell-type specificities of X-linked and autosomal genes, which showed a notable excess (12–60%) of X-linked genes with predominant expression in SG across all eutherians, in agreement with a previous mouse study[17] and includes conserved genes with key spermatogenic functions such as *TEX11* (Figs. 3a and 4b, Extended Data Fig. 9a and Supplementary Table 10). The higher-resolution primate data showed that the X chromosome is also enriched for genes expressed in leptotene SC (Fig. 4b), presumably reflecting transcript carry-over from SG, given the global transcriptional silencing of the genome during the leptotene stage[41] (Supplementary Fig. 1b and Supplementary Table 3) and the observation that most X-linked genes expressed in leptotene SC are also expressed in differentiated SG (Supplementary Fig. 4). Notably, we detected enrichments of genes with SG-specific expression also on the opossum and platypus X chromosomes, thus revealing this to be a shared pattern across mammals (Fig. 4b and Extended Data Fig. 9a,b).

We also uncovered an enrichment of genes with predominant expression in Sertoli cells across X chromosomes (except for platypus, for which Sertoli cells could not be unambiguously distinguished) (Fig. 4b). Altogether, our observations suggest that sex-related forces favoured the independent accumulation of SG- and Sertoli-specific genes on the X during the

evolution of the different sex-chromosome systems. Consistently, autosomes in outgroup species corresponding to the different mammalian X chromosomes (for example, platypus chromosome 6, which is homologous to the therian X) do not show any excess of SG- or Sertoli-expressed genes (Extended Data Fig. 9a,c), which means that the ancestral autosomes that gave rise to present-day sex chromosomes were not enriched for such genes. Overall, we propose that the accumulation of SG- and Sertoli-specific genes was facilitated by the specific selective environment on the X chromosome, in which male-beneficial mutations are always visible to selection because of the single-copy (hemizygous) status of the X chromosome in males[3,17,21].

We next sought to separate X- and Y-bearing spermatids to investigate their distinct transcriptomal properties during spermiogenesis. Such an analysis is probably not possible using single-whole-cell transcriptomic data because X and Y spermatids remain connected by cytoplasmic bridges and hence are thought to contain similar cytoplasmic transcript pools[42]. However, our single-nucleus data should afford the separation of X and Y spermatids. Indeed, on the basis of differential X and Y transcript contents (Methods), we were able to separate spermatids into distinct X and Y lineages across mammals (Fig. 4c and Supplementary Table 3). As expected, our approach failed to separate X and Y spermatids in available human[23] and mouse[43] scRNA-seq datasets (Extended Data Fig. 10), supporting the notion of substantial transcript exchange across X- and Y-spermatid cells through cytoplasmic bridges[42], although this equilibration may not be complete[44].

A differential expression analysis between X and Y spermatids identified, as expected, most sex-chromosome genes, including gametologues (that is, genes with homologous counterparts on X and Y chromosomes), such as the translational regulatory genes *DDX3X*/*DDX3Y* (Extended Data Fig. 11 and Supplementary Table 11). However, we also found autosomal genes, especially in some species (for example, human), which might reflect trans-regulatory effects associated with the X and Y chromosomes (Extended Data Fig. 11, Supplementary Table 11 and Methods).

We then assessed gene expression across spermatogenesis separately for X- and Y-linked genes. We traced a substantial dip in X transcript abundances around the pachytene stage of meiosis across therians (Fig. 4d), which

reflects the process of MSCI[13], a sex-chromosome-specific instance of the general epigenetic phenomenon of meiotic silencing of unsynapsed chromatin[13] (MSUC). An analysis of Y transcripts provides consistent results but at lower resolution owing to the very small number of Y-linked genes (Extended Data Fig. 12a). Previous work did not find evidence for MSCI in monotremes[45], suggesting that MSCI originated in the therian ancestor after the separation from the monotreme lineage[45]. We revisited this question in platypus using a new assembly and annotation[46] that includes a detailed definition of sexually differentiated regions (SDRs) versus pseudoautosomal regions (PARs), which are large in monotremes (Fig. 4a). We proposed that expression signals from the large PARs, which are expected to synapse and hence not to be affected by MSUC, might have prevented the detection of MSCI in previous studies.

Indeed, whereas the joint analysis of all platypus X-linked genes only shows a small expression dip around the pachytene stage (Fig. 4d, upper platypus graph), an analysis only of SDR genes reveals a strong reduction of X transcript levels. By contrast, PAR genes show stable expression levels across spermatogenesis (Fig. 4d, lower graph). Moreover, the difference in transcript abundances between SDRs and PARs owing to MSCI is visible for all five platypus X chromosomes (Extended Data Fig. 12b). Notably, our assessment of the completeness of MSCI across species reveals that platypus MSCI is as complete as that in other species for which there is little or no MSCI escape[23,47,48] (Extended Data Fig. 12c). The presence of MSCI at the SDRs in monotremes is consistent with the partial association of platypus sex chromosomes with perinucleolar repressive histone modifications at the pachytene stage[45]. Altogether, our data reveal that efficient MSCI is common to all mammalian sex-chromosome systems, which indicates that the general mechanism of MSUC is an ancestral mammalian feature.

Discussion

Our analyses uncovered that the previously observed rapid evolution of the testis[1,2,3,4,5,6,7] is driven by an accelerated rate of molecular innovation in late spermatogenesis, in particular in rSD. Our findings suggest a scenario in which the accelerated fixation of regulatory changes, amino acid altering

substitutions and new genes during evolution in late spermatogenesis (presumably driven by sperm competition) was facilitated by reduced pleiotropic constraints, a transcriptionally permissive chromatin environment and potentially haploid selection for spermatid genes that are not rendered effectively diploid through transcript exchange through cytoplasmic bridges. In agreement with late spermatogenic stages explaining the rapid evolution of the testis, early spermatogenic cell types and somatic cells show patterns of constraint and innovation similar to those of cell types in the brain (Extended Data Fig. 3a,b,d), an organ that evolves slowly at the molecular level[1,2,3]. We note that differences in cell-type abundances across mammals, which are pronounced for the testis[7] (Extended Data Fig. 1f), presumably also contributed to the rapid gene expression divergence of this organ[6].

Our cross-species comparisons of individual genes revealed temporal expression differences across species, which were probably facilitated by reduced pleiotropic constraints. Our results thus provide an extensive list of candidates whose contributions to the evolution of species-specific spermatogenesis phenotypes can be experimentally scrutinized. We also uncovered conserved expression programs underlying spermatogenic processes ancestral to individual mammalian lineages and mammals as a whole.

Further analyses illuminated the role of sex chromosomal genes in spermatogenesis. We found that genes predominantly expressed in SG and Sertoli cells independently accumulated on X chromosomes across mammals and their two sex-chromosome systems during evolution. This suggests that X chromosomes have been shaped by strong male-related selective forces[21], leading to the emergence of X-linked genes with functional roles in testis cell types in which active transcription is possible. Indeed, in addition to the SG and Sertoli-cell enrichment found here, previous work showed that these selective pressures led to the repeated duplication of genes on the X that facilitated expression after meiosis[15].

Our ability to separate X- and Y-bearing spermatids and the availability of a new platypus genome assembly[46] unveiled that MSCI is a general feature of mammalian sex-chromosome systems, indicating that MSUC was already

present in the common mammalian ancestor. Previous work did not find evidence for MSCI in birds[49], which raises the question whether MSUC arose in the common mammalian ancestor after the split from the reptile lineage roughly 310 million years ago or whether this mechanism was lost in the avian lineage and arose earlier in evolution. The latter scenario would be consistent with the observation of MSCI in invertebrate species[13].

Our data and results, together with the accompanying online resource we developed (https://apps.kaessmannlab.org/SpermEvol/) to facilitate the exploration of our data, provide an extensive resource for investigating the biology of the testis and associated fertility disorders across mammals. Future studies should seek to complement our snRNA-seq data to overcome its limitations. scRNA-seq data will be valuable for inferring transcriptome patterns unique to the cytoplasm, single-cell full-length transcript data are needed to assess the pronounced isoform diversity of the testis[11], and single-cell translatome data[50] are required to understand the contribution of post-transcriptional changes[6] to the evolution of spermatogenesis.

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Biological samples and ethics statement

We generated snRNA-seq data for adult testis samples from human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), gorilla (*Gorilla gorilla*), lar gibbon (*Hylobates lar*), rhesus macaque (*Macaca mulatta*), common marmoset (*Callithrix jacchus*), mouse (*Mus musculus*, strain: RjOrl:SWISS; Janvier Laboratories), grey short-tailed opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*) and chicken (red jungle fowl, *Gallus gallus*) (Supplementary Table 2). In addition, we produced bulk RNA-seq data for chimpanzee, gorilla, gibbon and marmoset

from the same individuals. Adult human testis samples used for in situ hybridization experiments were obtained from orchiectomy specimens from three individuals with testicular cancer. Tissue adjacent to the tumour that was devoid of cancer cells and germ cell neoplasia in situ, and tubules with normal spermatogenesis were used. Other adult primate testis tissue was obtained from a western chimpanzee and a Bornean orangutan (*Pongo pygmaeus*).

Our study complies with all relevant ethical regulations with respect to both human and other species' samples. Human samples underlying the snRNA-seq data were obtained from scientific tissue banks (http://medschool.umaryland.edu/btbank/) or dedicated companies (https://www.tissue-solutions.com/); informed consent was obtained by these sources from donors before death or from next of kin. The samples used for the RNA in situ hybridization experiments were obtained from the tissue biobank at the Department of Growth and Reproduction (Rigshospitalet, Copenhagen, Denmark) containing orchiectomy specimens from individuals with testicular cancer (The Danish Data Protection Agency, permit number J. no. 2001-54-0906). All patients have given informed consent for donating the residual tissues for research. The use of all human samples for the type of work described in this study was approved by an ethics screening panel from the European Research Council (ERC) and local ethics committees: from the Cantonal Ethics Commission in Lausanne (authorization 504/12); from the Ethics Commission of the Medical Faculty of Heidelberg University (authorization S-220/2017) and from the regional medical research ethics committee of the capital region of Copenhagen (H-16019637). All primates used in this study suffered sudden deaths for reasons other than their participation in this study and without any relation to the organ sampled. The use of all other mammalian samples for the type of work described in this study was approved by ERC ethics screening panels.

Nuclei isolation

For the samples from therian species we developed a nuclei preparation method that includes fixation with dithio-*bis*(succinimidyl propionate) (DSP; or Lomant's Reagent), a reversible cross-linker that stabilizes the isolated nuclei. The method was adapted from protocols used for the fixation

of single-cell suspensions[51] and for the isolation of single nuclei from archived frozen brain samples[52]. Tissue pieces weighing roughly 5 mg were homogenized in 100–150 µl mg^{-1} of prechilled lysis buffer (250 mM sucrose, 25 mM KCl, 5 mM $MgCl_2$, 10 mM HEPES pH 8, 1% BSA, 0.1% IGEPAL and freshly added 1 µM DTT, 0.4 U µl^{-1} RNase Inhibitor (New England BioLabs), 0.2 U µl^{-1} SUPERasIn (ThermoFischer Scientific)) and lysed for 5 min on ice. The lysate was centrifuged at 100*g* for 1 min at 4 °C. The supernatant was transferred to a new reaction tube and centrifuged at 500*g* for 5 min at 4 °C. The supernatant was removed and the pellet resuspended in 0.67 vol. (of volume lysis buffer used) of freshly made fixation solution (1 mg ml^{-1} DSP in PBS) and incubated for 30 min at room temperature. The fixation was quenched by addition of Tris-HCl to a final concentration of 20 mM. The fixed nuclei were pelleted at 500*g* for 5 min at 4 °C. The supernatant was removed and the pellet resuspended in 0.67 vol. of Wash Buffer (250 mM sucrose, 25 mM KCl, 5 mM $MgCl_2$, 10 mM Tris-HCl pH 8, 1% BSA and freshly added 1 µM DTT, 0.4 U µl^{-1} RNase Inhibitor, 0.2 U µl^{-1} SUPERasIn). This was centrifuged at 500*g* for 5 min at 4 °C. The supernatant was removed and the pellet resuspended in 0.5 vol. of PBS. Then nuclei were strained using 40 µm Flowmi strainers (Sigma). For estimation of nuclei concentration, Hoechst DNA dye was added and the nuclei were counted using Countess II FL Automated Cell Counter (ThermoFischer Scientific).

For platypus and chicken, a similar preparation method was used, but the nuclei were not fixed, given that this protocol gave optimal results for these species (the fixation protocol failed to yield data of adequate quality). In brief, tissue pieces weighing roughly 5 mg were homogenized in 100–150 µl mg^{-1} of prechilled lysis buffer (250 mM sucrose, 25 mM KCl, 5 mM $MgCl_2$, 10 mM Tris-HCl pH 8, 1% BSA, 0.1% IGEPAL and freshly added 1 µM DTT, 0.4 U µl^{-1} RNase Inhibitor, 0.2 U µl^{-1} SUPERasin) and lysed for 5 min on ice. The lysate was centrifuged at 100*g* for 1 min at 4 °C. The supernatant was transferred to a new reaction tube and centrifuged at 500*g* for 5 min at 4 °C. The supernatant was removed and the pellet resuspended in 0.67 vol. of Wash Buffer. This was centrifuged at 500*g* for 5 min at 4 °C. The supernatant was removed and the pellet resuspended in 0.5 vol. of PBS. Then nuclei were strained using 40-µm Flowmi strainers (Sigma). For

estimation of nuclei concentration, Hoechst DNA dye was added and the nuclei were counted using Countess II FL Automated Cell Counter (ThermoFischer Scientific).

Preparation of sequencing libraries

For the construction of snRNA-seq libraries, Chromium Single Cell 3′ Reagent Kits (10X Genomics; v.2 chemistry for human, chimpanzee, bonobo, gorilla, gibbon, macaque, marmoset, mouse and opossum; v.3 chemistry for platypus and chicken) were used according to the manufacturer's instructions. Then 15,000 to 20,000 nuclei were loaded per lane in the Chromium microfluidic chips and complementary DNA was amplified in 12 PCR cycles. Sequencing was performed with NextSeq550 (Illumina) according to the manufacturer's instructions using the NextSeq 500/550 High Output Kit v.2.5 (75 cycles) with paired-end sequencing (read lengths of read1 26 bp, read2 57 bp; index1 8 bp, roughly 170 to 380 million reads per library for v.2 chemistry; read lengths of read1 28 bp, read2 56 bp Index1 8 bp, roughly 247 to 306 million reads per library for v.3 chemistry) (Supplementary Table 2).

For bulk RNA-seq data generation, RNA was extracted using the RNeasy Micro kit (QIAGEN). The tissues were homogenized in RLT buffer supplemented with 40 mM DTT. The RNA-seq libraries were constructed using the TruSeq Stranded messenger RNA LT Sample Prep Kit (Illumina) as described in ref. [2]. Libraries were sequenced on Illumina NextSeq550 using a single-end run (read1 159 bp; index1 7 bp) with roughly 24–60 million reads per library (Supplementary Table 2).

Genome and transcript isoform annotation

Given that the quality of genome annotation differs substantially between the studied species and given the specific and widespread transcription of the genome in the testis[11], we refined and extended previous annotations from Ensembl[53] on the basis of testis RNA-seq data. Specifically, akin to the procedure described in refs. [2,6], we used previous extensive stranded poly(A)-selected RNA-seq datasets[2,54] (100 nt, single-end) for human, macaque, mouse, opossum, platypus and chicken, and generated and used

stranded poly(A)-selected RNA-seq datasets (159 nt, single-end) for chimpanzee, gorilla, gibbon and marmoset. For each species, we downloaded the reference genome from Ensembl release 87 (ref. [53]): hg38 (human), CHIMP2.1.4 (chimpanzee), rheMac8 (rhesus macaque), C_jacchus3.2.1 (marmoset), mm10 (mouse), monDom5 (opossum), and galGal5 (chicken); from Ensembl release 96 (ref. [55]): gorGor4 (gorilla) and Nleu_3.0 (gibbon); and from Ensembl release 100 (ref. [56]): mOrnAna1.p.v1 (platypus). Raw reads were first trimmed with cutadapt v.1.8.3 (ref. [57]) to remove adapter sequences and low-quality (Phred score <20) nucleotides, then reads shorter than 50 nt were filtered out (parameters: --adapter=AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC --match-read-wildcards --minimum-length=50 -q 20). Processed reads were then mapped to the reference transcriptome and genome using Tophat2 v.2.1.1 (ref. [58]) (parameters: --bowtie1 --read-mismatches 6 --read-gap-length 6 --read-edit-dist 6 --read-realign-edit-dist 0 --segment-length 50 --min-intron-length 50 --library-type fr-firststrand --max-insertion-length 6 --max-deletion-length 6). Next, we assembled models of transcripts expressed using StringTie v.1.3.3 (ref. [59]) (parameters: -f 0.1 -m 200 -a 10 -j 3 -c 0.1 -v -g 10 -M 0.5). Stringent requirements on the number of reads supporting a junction (-j 3), minimum gap between alignments to be considered as a new transcript (-g 10) and fraction covered by multi-hit reads (-M 0.5) were used to avoid merging independent transcripts and to reduce the noise caused by unspliced or incompletely spliced transcripts. We compared the assembled transcript models to the corresponding reference Ensembl annotations using the cuffcompare program v.2.2.1 from the cufflinks package[60]. Finally, we combined the newly identified transcripts with the respective Ensembl gene annotation into a single gtf file. We extended the original Ensembl annotations by 2–61 Mbp with new transcripts and by 23–49 Mbp with new splice isoforms (Supplementary Table 1).

Raw reads processing

CellRanger v.3.0.2 was used for platypus and chicken, and CellRanger v.2.1.1 for the other species in line with the used Chromium chemistry. The CellRanger mkref function was used with default settings to build each species reference from genomic sequences and customized extended annotation files (Supplementary Table 1). Given that pre-mRNA transcripts

are abundant in nuclei[61], exons and introns features were concatenated as described in the CellRanger v.2.1.1 documentation for considering intronic and exonic reads for gene expression quantification. The CellRanger count function was used with default settings to correct droplet barcodes for sequencing errors, align reads to the genome and count the number of UMIs for every gene and barcode combination.

Identification of usable nuclei

We used a combined approach for detection of usable nuclei. This was done to optimally account for the lower RNA content of nuclei compared to the whole cells. Specifically, to identify usable nuclei, we used a knee point-based approach combined with the fraction of intronic reads as a marker of pre-mRNA transcripts (abundant in the nucleus) (Supplementary Fig. 1f) and *MALAT1* (nuclear-enriched long non-coding (lnc)RNA) expression as a marker of nuclei (when present in the genome of a given species).

Quality control of filtered cells

For each sample independently, high-quality nuclei were selected removing outliers on the basis of the number of UMIs and the percentage of mitochondrial RNA (Supplementary Fig. 1b,c,e). We created a Seurat[62] object using the Seurat R package v.3.1.4 from the subset raw UMI count table generated by CellRanger corresponding to the usable droplets identified upstream, normalized the data using the NormalizeData function, identified the top 10,000 most variables genes using the FindVariableFeatures function, scaled the data using the ScaleData function, performed the PCA using the RunPCA function and calculated the Louvain clusters using the FindNeighbors (parameters dims = 1:20) and FindClusters (dims = 1:20, resolution = 0.5) functions. To optimally account for the fact that testis cell types have diverse transcriptome characteristics[11], we filtered out outlier droplets for each cluster independently with values lower than the first quartile (Q1) − 1.5 × IQR (interquartile range) and higher than the third quartile (Q3) + 1.5 × IQR for both the UMI content and the fraction of mitochondrial RNA. Then, we removed potential doublets using doubletFinder_v3 function of DoubletFinder[63] v.2.0.1 (parameters PCs =

1:20, pN = 0.25, nExp = 5% of the total number of cells, identifying pk using paramSweep_v3, summarizeSweep and find.pK functions).

Integration of datasets

From the previously filtered UMI count tables, we created Seurat objects for every sample independently, normalized the data and identified the top 10,000 most variable genes. Next, for each species independently, we applied the Seurat[62] anchoring approach using FindIntegrationAnchors and IntegrateData functions with 20 principal components to integrate all datasets together into a single Seurat object correcting for the batch effect. For each integrated species-specific Seurat object, we normalized (NormalizeData function) and scaled the data (ScaleData function) and performed a PCA (RunPCA function). Louvain clusters were calculated using FindNeighbors and FindClusters functions (parameters dims = 1:20, 1:20, 1:20, 1:20, 1:20, 1:20, 1:20, 1:17, 1:8, 1:10 and 1:10, and resolution = 0.5, 0.5, 0.2, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.3 and 0.5, respectively, for human, chimpanzee, bonobo, gorilla, gibbon, macaque, marmoset, mouse, opossum, platypus and chicken). The uniform manifold approximation and projection (UMAP) embedding coordinates were calculated using the RunUMAP function (parameters dims = 1:20, 1:20, 1:20, 1:20, 1:20, 1:20, 1:20, 1:17, 1:10, 1:10 and 1:10, and min_dist = 0.3, 0.3, 0.1, 0.1, 0.3, 0.3, 0.3, 0.1, 0.2, 0.3 and 0.6, respectively, for human, chimpanzee, bonobo, gorilla, gibbon, macaque, marmoset, mouse, opossum, platypus and chicken). We note that —consistent with the high correlation between biological replicates (Supplementary Fig. 5a,b)—the data already integrate well before the batch correction (Supplementary Fig. 18c,d). We also note that key marker genes are expressed in the same integrated areas across replicates when assessing their expression in the different replicates using the integrated object coordinates (Supplementary Fig. 18e), which supports that the integration is correct.

All primate datasets were merged using the LIGER[64] (v.0.5.0) integration tool. A LIGER object was created using the createLiger function based on primate 1:1 orthologues from Ensembl release 87, and normalized with normalize, selectGenes and scaleNotCenter functions with default settings. Then, the joint matrix was factorized using the optimizeALS function (k =

20) and the quantile normalization was performed with the quantile_norm and default settings. The Louvain clusters were calculated with the louvainCluster function and default settings as well as UMAP coordinates with the runUMAP function (n_neighbors = 100, min_dist = 0.2).

Estimation of expression levels and normalization

The gene UMI counts per cell were normalized using the Seurat R package and its NormalizeData function. Therefore, the UMI counts of each gene in each cell are divided by the total UMI counts of each cell, multiplied by 10,000 and log transformed.

Cell-type assignment

We identified the main cell-type populations from the primate integrated, mouse, opossum, platypus and chicken objects independently using known marker genes[65,66] mostly from human and mouse and their respective 1:1 orthologues in the other species. *CLU* marks Sertoli cells; *TAGLN* and *ACTA2* peritubular and smooth muscle cells; *CD34* and *TM4SF1* endothelial cells; *APOE* and *CD74* macrophages; *STAR* and *CYP11A1* Leydig cells; *GFRA1, PIWIL4* (undifferentiated), *DMRT1*(differentiated) and *STRA8* SG; *SYCE1* (leptotene), *SYCP1* (zygotene), *PIWIL1* (pachytene)*, SYCP2, TANK* and *AURKA* SC; *LRRIQ1* (early), *ACRV1* and *SPACA1* (late) rSD; and *SPATA3*, *NRBP1, PRM1* and *GABBR2* eSD. Cell-type assignment was robustly reinforced by complementary analyses and metrics such as UMAP coordinates, pseudotime trajectories, transcriptional activities (UMI counts) and previous knowledge.

Pseudotime

Pseudotime trajectories were calculated using the slingshot v.1.2.0 R package[67]. We applied the getLineages function with the upstream calculated clusters and UMAP embedding coordinates of the germ cells to obtain connections between adjacent clusters using a minimum spanning tree. We provided the starting and ending clusters on the basis of the previous cell-type assignment with known marker gene expression. Then we applied the getCurves function to the obtained lineages to construct smooth

curves and order the cells along a pseudotime trajectory. Pseudotime values are highly variable depending on used tools, thus we ordered the cells one by one on the basis of their pseudotime values and divided their rank by the total number of cells, to obtain evenly distributed values between 0 and 1. Finally, we validated the obtained pseudotime trajectories on the basis of previous cell-type assignments, expression patterns of marker genes and UMAP embedding coordinates.

Marker gene identification

To precisely identify marker genes along spermatogenesis, we grouped the germ cells into 20 evenly distributed bins along the pseudotime trajectory for each species. Then, we applied the FindAllMarkers function from the Seurat[62] R package (parameter only.pos = T, min.pct = 0.25, logfc.threshold = 0.25, return.thresh = 0.05) of the Seurat v.3.1.4 R package to the 22 groups (20 germline groups, the Sertoli and other somatic cell groups) (Extended Data Fig. 1 and Supplementary Table 4).

Phylogenetic trees

Phylogenetic trees and indicated divergence times (Figs. 1a and 4a) are based on TimeTree[68] (v.5) (http://www.timetree.org/).

Orthologous gene sets

We used four different sets of orthologous genes in our study: (1) comparative analyses involving all 11 amniote species were performed using 4,498 1:1 orthologue genes that are expressed (that is, one UMI in at least three cells of any cell type) across all species (among a total of 8,045 1:1 orthologues). (2) Comparative analyses involving the seven primate species were performed using 8,451 1:1 orthologue genes expressed across all primate species (among 11,948 1:1 orthologues). (3) The comparative Sertoli-germ cell communication analysis was based on mapping 35,186 human testis-expressed genes to 1:1 orthologous genes in the other species (macaque 13,090; mouse 14,302; opossum 10,865 and chicken 10,515). (4) Species-specific analyses were performed using all genes expressed in a given species (roughly 15,000 genes per cell type; Supplementary Fig. 1d).

Orthologous gene sets were extracted from Ensembl[53] using the biomaRt R package v.2.40.5.

Global patterns of gene expression differences across mammals

Pseudo-bulk samples were generated using the AverageExpression function of the Seurat R package with various groups of cells depending on the pseudo-bulk samples produced in the study. For the analyses presented in Fig. 1b, we performed the PCA of normalized expression in amniote testicular cell types (pseudo-bulks) for each individual based on 4,498 1:1 amniote orthologues. PCA was performed using the prcomp function of the stats R package. For Fig. 1c, we constructed gene expression trees (as described in ref. [1]) using the neighbour-joining approach, on the basis of pairwise expression distance matrices between corresponding pseudo-bulk samples for the different cell types across species. The distance between samples was computed as $1 - \rho$, where ρ is Spearman's correlation coefficient and was computed using the cor function of the stats R package. The neighbour-joining trees were constructed using the ape R package v.5.3. The reliability of branching patterns was assessed with bootstrap analyses (the 4,498 1:1 amniote orthologues were randomly sampled with replacement 1,000 times). The bootstrap values are the proportions of replicate trees that share the branching pattern of the majority-rule consensus tree shown in the figures (Fig. 1c and Extended Data Fig. 2a). The total tree length was calculated by removing the intra-species variability between individuals (Fig. 2a).

Evolutionary forces

In Fig. 2c, we plotted the median pLI score[69] across expressed genes ($\ge$1 UMI) in each nucleus. We obtained the pLI scores from ref. [70]. For Fig. 2d, we used a set of neutrally ascertained knockouts consisting in 4,742 protein-coding genes, 1,139 of which are classified as lethal. For each cell, the denominator is the number of genes expressed that were tested for lethality and the numerator the genes among those that resulted in a lethal phenotype. Tested genes for viability and associated phenotype information were downloaded from the International Mouse Phenotyping Consortium[25]. For Fig. 2e (and Extended Data Fig. 3a), we used the average d*N*/d*S* values

across 1:1 orthologues in primates. For each cell, the mean d*N*/d*S* value is plotted. Conserved 1:1 orthologues across six primates (human, chimpanzee, gorilla, gibbon, macaque and marmoset) as well as their coding and protein sequences were extracted from Ensembl[53], providing a set of 11,791 protein-coding genes. For each species and orthologue the longest transcript was extracted. Orthologous protein sequences were aligned using clustalo v.1.2.4; then pal2nal v14 was used (with protein sequences alignments and coding sequences as input) to produce codon-based alignments. The codeml software from the PAML package[71] v.4.9 was used to estimate d*N*/d*S* ratios. The M0 site model was applied to the orthologue alignments to estimate one average dN/dS ratio per orthologous gene set across species (parameter NSites = 0, model = 0). In Fig. 2f (and Extended Data Fig. 3b), we plotted the percentage of positively selected genes expressed across nuclei. For each nucleus, the denominator is the number of expressed genes that were tested for signatures of positive selection, and the numerator is the number of genes among those with evidence for positive selection. We used sets of genes previously identified as carrying evidence for coding-sequence adaptation in primates[72] (331 positively selected genes out of 11,170 genes tested) and mammals[73] (544 positively selected genes out of 16,419 genes tested).

In Fig. 2g (and Extended Data Fig. 3c), we plotted the average phylogenetic age of expressed genes across somatic and germ cells. The phylogenetic age of genes is an index that gives greater weight to young new genes (as described in ref. [2,74]). The range of the score differs between species depending on the number of outgroup lineages available (more lineages allowed for more details in the phylogeny) and therefore this index cannot be compared across species, only within species (that is, across cells and cell types). The phylogenetic age of genes was obtained from GenTree (http://gentree.ioz.ac.cn/) with Ensembl release 69 (ref. [74]). In Fig. 2h (and Extended Data Fig. 3d), we plotted the percentage of the cell transcripts originating from protein-coding genes and intergenic elements. Gene biotypes were obtained from Ensembl. Intergenic elements are all elements that are not protein-coding genes (lncRNAs, pseudogenes, pseudogenes and other sequences). In Fig. 2i (and Extended Data Fig. 3e), normalized $\log_2$-transformed median expression values across replicates at the transcriptome (e^{tr}) and translatome (e^{tl}) layers were used to calculate translation efficiency

(TE = $\log_2(e^{tr}) - \log_2(e^{tl})$) in testis (as described in ref. [6]) from RNA-seq and Ribo-seq data[6], respectively. Translation efficiency values were calculated at the whole testis level, thus only cell-type-specific genes (for which 60% of their transcripts at the whole testis level are concentrated in a single cell type) were used. Higher values show a more efficient translation of transcripts, whereas lower values indicate relative translational repression. For Fig. 2j (and Extended Data Fig. 3f,g), we used time- and tissue-specificity indexes of expressed genes across somatic and germ cells in testis. As described in ref. [2], tissue and time specificity indexes are calculated from RNA-seq data across organs and developmental stages. Both indexes range from 0 (broad expression) to 1 (restricted expression). The indexes were obtained from ref. [2]. For each nucleus, we plotted the median index across expressed genes. In Fig. 2k, we plotted the percentage of genes causing infertility when knocked out (out of 3,252 knockouts, 173 of which caused infertility). Tested genes for infertility and associated phenotype information were downloaded from the International Mouse Phenotyping Consortium database[25]. For each nucleus, the denominator corresponds to the number of genes expressed that were tested for infertility and the numerator to the genes among those that resulted in an infertility phenotype.

Gene expression trajectories along spermatogenesis

We compared gene expression trajectories along spermatogenesis across primates using human, chimpanzee, bonobo, gorilla, gibbon, macaque and marmoset (based on Ensembl 87 orthologues), and across amniotes using human (as a representative of primates), mouse, opossum, platypus and chicken (based on Ensembl 100 orthologues). To compare robustly expressed genes, we used genes that are expressed in at least 5% of the cells in at least one cluster in all considered species. We used the mfuzz package[75] (v.2.44.0), an unsupervised soft clustering method, to cluster gene expression trajectories along spermatogenesis (eight cell types in primates; four cell types in amniotes) across species using Dmin and mestimate functions to estimate the number of clusters and the fuzzification parameter (Supplementary Fig. 2). As described in ref. [2], we inferred within a phylogenetic framework the probability that there were changes in trajectories along spermatogenesis, that is, that genes changed their cluster assignment in specific branches, using a 5% probability cut-off.

Trajectory conservation score

We calculated a global trajectory conservation score across species for each 1:1 orthologous gene set. For a given orthologous gene set, this score corresponds to the log-transformed probability that all members fall into the same mfuzz trajectory cluster as:

$${\rm{Conservation}}_{{\rm{score}}_{g}}={\log }_{2}({\sum }_{i\in c}{\prod }_{j\in s}{P}_{g,i,j})$$

where *g* corresponds to a given orthologous gene set, *c* to all mfuzz trajectory clusters (1–9 for primates, 1–12 for amniotes), *s* to all species (human, chimpanzee, bonobo, gorilla, gibbon, macaque and marmoset for primates; human, mouse, opossum, platypus and chicken for amniotes) and $P_{g,i,j}$ to the probability that the gene *g* of the species *j* falls into the cluster *i*. A higher conservation score indicates a greater global trajectory conservation. As a proof of concept, we plotted the trajectory conservation score for conserved and changed trajectories that revealed a significant higher conservation score for conserved trajectories (Extended Data Fig. 5a).

RNA in situ hybridization and expression quantification

Fresh testicular tissue was fixed in GR-fixative (7.4% formaldehyde, 4% acetic acid, 2% methanol, 0.57% sodium phosphate, dibasic and 0.11% potassium phosphate, monobasic) overnight (for at least 16 h) at 4 °C, dehydrated and embedded in paraffin. The in situ hybridization experiments were carried out on 4 µm sections mounted on SuperFrost Plus Slides (ThermoFisher Scientific) using the RNAScope 2.5 HD Detection Reagent RED kit according to the manufacturer’s recommendations (Advanced Cell Diagnostics). Briefly, testicular tissue sections were dewaxed in xylene and washed in 100% ethanol followed by treatment with hydrogen peroxide for 10 min. Target retrieval was performed for 15 or 30 min (see Supplementary Table 6 for specifications for each probe and species) using a steamer, followed by treatment with protease plus for 30 min at 40 °C. The slides were hybridized with the target probe (Supplementary Table 6) for 2 h at 40 °C followed by a series of signal amplifications (with amplification around 5 for 30 or 60 min). The sections were counterstained with Mayer’s

haematoxylin and mounted with Vectamount Permanent Mounting Medium (Vector Laboratories). The negative control probe *DapB* (a bacterial RNA) was run in parallel with the target probes and showed ≤5% positive cells in each section.

For *ADAMTS17* and *MYO3B*, positive (that is, with red dots) rSD and SG were counted. For each section (human $n = 3$, chimpanzee $n = 1$, orangutan $n = 1$), ten tubules were counted using the NDP.viewPlus software (Hamamatsu Photonics). Two independent observers (S.B.W. and K.A.) counted positive and negative rSD and SG. No discrimination in intensity of the dots or the number of dots per cell was performed. Cell-type identification was performed on the basis of nucleus morphology and localization in the tubule. Only SG lining the edge of the tubules were counted (Supplementary Fig. 3a). The inter-observer variance was found to be 7 and 9% for rSD and SG, respectively. For *RUBCNL*, quantification of staining intensity was performed with R v.3.6.1 using countcolors[76] (v.0.9.1) and colordistance[77] (v.1.1.1) packages. For each section (human $n = 3$, chimpanzee $n = 1$, orangutan $n = 1$), ten tubules were divided into three parts: area dominated by SG, area dominated by SCs (also containing Sertoli cells) and area dominated by spermatids (no distinction between the different types of spermatid) (Supplementary Fig. 3b). In each tubular area, the number of cells was counted manually using the NDP.view2Plus software (v.2.8.24). Then, the pixels occupied by red staining were quantified and the expression level for each cell type was calculated by dividing the stained pixels by the number of cells. For each picture, the stained pixels for each cell type were normalized by the total amount of stained pixels.

Gene Ontology analysis

Enriched terms in the Gene Ontology[32] analyses of genes with conserved and diverged expression trajectories were identified using the goana function of the limma R package, v.3.40.6 (default parameters).

Sertoli-germ cell communication analysis

We identified ligand–receptor interactions underlying Sertoli-germ cell communications for human, macaque, mouse, opossum and chicken, respectively, using the CellPhoneDB[78] (v.2) approach and recommended parameter settings (parameters method statistical_analysis). To apply CellPhoneDB, which uses a database of human ligand–receptor interactions, to the data for the other species, we mapped human testis-expressed genes to their corresponding 1:1 orthologues in each of these species. Enriched receptor–ligand interactions between two cell types are predicted on the basis of expression of a receptor by one cell type and a ligand by another cell type[78]. Only receptors and ligands expressed in more than 10% of the cells in each cell type were considered for pairwise comparisons between all cell types in the dataset. CellPhoneDB uses empirical shuffling to calculate which ligand–receptor pairs show significant ($P < 0.05$) cell-type specificity[78]. Significant interactions across species were illustrated using the R package UpSetR (v.1.4.0). Finally, given potential false positive (and negative) predictions of CellPhoneDB and similar approaches[79], we consider interactions that are predicted for several species and probably reflect evolutionary conservation (Extended Data Fig. 8 and Supplementary Table 9), such as those reported in the main text, to be more reliable than species-specific predictions.

Cell-type and testis-specific genes per chromosome

Testis-specific genes were obtained from previously generated RNA-seq data[2] of adult organs (RPKM $\ge$ 1 in testis and RPKM < 1 in brain, cerebellum, heart, kidney and liver). Among these, cell-type-specific genes were studied for each chromosome. Genes with predominant expression in specific somatic cell types were identified using the FindAllMarkers function (parameter only.pos = TRUE, min.pct = 0.05, logfc.threshold = 0.25, return.thresh = 0.05). Predominant expression of genes in specific germ cell types was assigned on the basis of the trajectory analyses (above); that is, predominant expression was assigned on the basis of the cell type in which the expression level of the gene peaks in the trajectory analysis. We then first calculated the percentage of genes located on a given chromosome among all genes in the genome (*x* axis of plots in Extended Data Fig. 9a,c; red horizontal line for X-linked genes in Fig. 4b). We then contrasted this with the percentage of testis-specific genes with predominant expression in a given cell type (*y* axis of plots in Extended Data Fig. 9a,c; *y* axis of plots in Fig. 4b for X-linked genes). Finally, the percentage of testis-specific genes per cell type and chromosome was statistically compared to the percentage of genes per chromosome in the genome using exact binomial tests.

Classification of X- and Y-bearing spermatids

The Y chromosome carries a low number of genes and is missing in some genome assemblies. Thus, we focused on the fraction of X transcripts in spermatids to classify them as X- or Y-bearing cells. For this, we fitted a Gaussian Mixture Model to the data with two components (bimodal distribution) independently for each replicate, using the function normalmixEM of the mixtools (v.1.2.0) R package. The two obtained normal distributions were used to classify X- (higher levels of X transcripts) and Y-bearing (lower levels of X transcripts) spermatids using 95% confidence intervals. Outlier and overlapping cells were not assigned to either category. Finally, we checked that the fraction of Y transcripts was significantly higher in Y-bearing spermatids (Fig. 4c). Bifurcating UMAPs (Fig. 4c) were obtained using X- and Y-linked genes in addition to

previously identified highly variable genes to perform the PCA associated with the UMAP coordinate calculation. For platypus, X transcripts are separated according to their location on the X chromosomes (that is, PARs and SDRs, respectively, as annotated in a previous study[46]). In the platypus genome assembly used[46], X and Y PARs are both assigned to the X chromosome. Thus, reported X transcripts may stem from X SDRs, X PARs or Y PARs, whereas reported Y transcripts only stem from Y SDRs. Illustrations of human and platypus sex chromosomes (with their respective PARs/SDRs) in Fig. 4a are based on previous work[46,80].

Transcriptomal differences between X and Y spermatids

We identified differentially expressed genes between X- and Y-spermatid populations using the FindMarkers function from Seurat[62] R package (parameters, default). A Wilcoxon rank-sum test was used to calculate *P* values that were adjusted using Bonferroni corrections for several tests. Only genes that were detected as expressed in at least 10% of cells from either of the two populations were tested. Genes that show, on average, at least a 0.25-fold higher expression (log_2-scale) in one of the populations, that are at the same time expressed in twice the number of cells in that population and that have an adjusted *P* value below 0.01 were considered to be differentially expressed. We note that several of the differentially expressed genes, including the most significant cases in human (Extended Data Fig. 11a), are putatively non-coding, which is noteworthy because lncRNAs are typically nuclear[81] and hence their differential expression levels are unlikely to be offset by transcript exchange between spermatid cells through cytoplasmic bridges. The three most Y-spermatid-specific transcripts are lncRNAs emanating from homologous low copy repeats (that is, segmental duplications) on chromosomes 13 (FAM230C), 21 (XLOC-095504) and 22 (FAM230F) that cause genomic disorders by triggering non-allelic homologous recombination events. They include the FAM230F lncRNA in the q11.2 low copy repeat region on chromosome 22 (22q11.2) that is particularly susceptible to non-allelic homologous recombination-generated deletions that lead to various congenital malformation disorders, including the DiGeorge syndrome, the most frequent microdeletion disorder[82].

MSCI completeness analysis

To identify potential MSCI escapee genes, we screened for X-linked genes with a significant increase in transcript abundance from SG to SC stages subject to MSCI, which ensures that potential escape genes are indeed actively transcribed in SC (that is, they do not merely represent genes expressed in SG with stable transcripts still detectable in SC), akin to previous work[23]. Specifically, we identified differentially expressed genes between SC and SG using the FindMarkers function from Seurat[62] R package (parameters, default). A Wilcoxon rank-sum test was used to calculate *P* values, which were then adjusted using a Bonferroni correction for several tests. Only genes that were detected as expressed in at least 10% of cells from either of the two cell-type populations were tested. Genes showing, on average, at least a 0.25-fold expression difference ($\log_e$-scale) between the two groups, and an adjusted *P* value below 0.05 were considered to be differentially expressed. X-linked genes in SDRs with significantly higher expression in SC than SG were considered to be potential escapees (Extended Data Fig. 12c).

General statistics and plots

Unless otherwise stated, all statistical analyses and plots were done in R v.3.6.2 (ref. [83]). Plots were created using ggplot2 v.3.2.1, tidyverse v.1.3.0, dplyr v.0.8.5, cowplot v.1.0.0 and pheatmap v.1.0.12.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw and processed bulk and snRNA-seq data have been deposited in ArrayExpress with the accession codes E-MTAB-11063 (human snRNA-seq), E-MTAB-11064 (chimpanzee snRNA-seq), E-MTAB-11067 (bonobo snRNA-seq), E-MTAB-11065 (gorilla snRNA-seq), E-MTAB-11066

(gibbon snRNA-seq), E-MTAB-11068 (macaque snRNA-seq), E-MTAB-11069 (marmoset snRNA-seq), E-MTAB-11071 (mouse snRNA-seq), E-MTAB-11072 (opossum snRNA-seq), E-MTAB-11070 (platypus snRNA-seq), E-MTAB-11073 (chicken snRNA-seq) and E-MTAB-11074 (chimpanzee, gorilla, gibbon and marmoset bulk RNA-seq) (https://www.ebi.ac.uk/arrayexpress/). All other data are available as Supplementary Information or available upon request. The testis gene expression at the single-nucleus level across the 11 studied species can be visualized using the shiny app we developed: https://apps.kaessmannlab.org/SpermEvol/.

Code availability

Custom scripts used to generate the results reported in the paper and processed data are available at https://github.com/evo-bio/Spermatogenesis.

References

1. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).

2. Cardoso-Moreira, M. et al. Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).

3. Necsulea, A. & Kaessmann, H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* **15**, 734–748 (2014).

4. Necsulea, A. et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).

5. Sarropoulos, i, Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510–514 (2019).

6. Wang, Z. Y. et al. Transcriptome and translatome co-evolution in mammals. *Nature* **588**, 642–647 (2020).

7. Ramm, S. A., Scharer, L., Ehmcke, J. & Wistuba, J. Sperm competition and the evolution of spermatogenesis. *Mol. Hum. Reprod.* **20**, 1169–1179 (2014).

8. Nielsen, R. et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).

9. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).

10. Kimmins, S. & Sassone-Corsi, P. Chromatin remodelling and epigenetic features of germ cells. *Nature* **434**, 583–589 (2005).

11. Soumillon, M. et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–2190 (2013).

12. Mazin, P. V., Khaitovich, P., Cardoso-Moreira, M. & Kaessmann, H. Alternative splicing during mammalian organ development. *Nat. Genet.* **53**, 925–934 (2021).

13. Turner, J. M. Meiotic silencing in mammals. *Annu. Rev. Genet.* **49**, 395–412 (2015).

14. Potrzebowski, L. et al. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.* **6**, e80 (2008).

15. Mueller, J. L. et al. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nat. Genet.* **40**, 794–799 (2008).

16. Mueller, J. L. et al. Independent specialization of the human and mouse X chromosomes for the male germ line. *Nat. Genet.* **45**, 1083–1087 (2013).

17. Wang, P. J., McCarrey, J. R., Yang, F. & Page, D. C. An abundance of X-linked genes expressed in spermatogonia. *Nat. Genet.* **27**, 422–426

(2001).

18. Julien, P. et al. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol.* **10**, e1001328 (2012).

19. Meunier, J. et al. Birth and expression evolution of mammalian microRNA genes. *Genome Res.* **23**, 34–45 (2013).

20. Zhang, Y. E., Vibranovski, M. D., Landback, P., Marais, G. A. & Long, M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* **8**, e1000494 (2010).

21. Rice, W. R. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38**, 735–742 (1984).

22. Lau, X., Munusamy, P., Ng, M. J. & Sangrithi, M. Single-cell RNA sequencing of the cynomolgus macaque testis reveals conserved transcriptional profiles during mammalian spermatogenesis. *Dev. Cell* **54**, 548–566.e7 (2020).

23. Shami, A. et al. Single-cell RNA sequencing of human, macaque, and mouse testes uncovers conserved and divergent features of mammalian spermatogenesis. *Dev. Cell* **54**, 529–547 e12 (2020).

24. Dickinson, M. E. et al. High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).

25. Koscielny, G. et al. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.* **42**, 802–809 (2014).

26. Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol.* **3**, 1159–1166 (2005).

27. Immler, S. Haploid selection in “diploid” organisms. *Annu. Rev. Ecol. Evol. Syst.* **50**, 219–236 (2019).

28. Cheng, X. & Sun, Q. RUBCNL/Pacer and RUBCN/Rubicon in regulation of autolysosome formation and lipid metabolism. *Autophagy* **15**, 1120–1121 (2019).

29. Zhu, Y. et al. Autophagy in male reproduction. *Syst. Biol. Reprod. Med.* **65**, 265–272 (2019).

30. Russell, D. L., Brown, H. M. & Dunning, K. R. ADAMTS proteases in fertility. *Matrix Biol.* **44–46**, 54–63 (2015).

31. Zhang, T., Oatley, J., Bardwell, V. J. & Zarkower, D. DMRT1 is required for mouse spermatogonial stem cell maintenance and replenishment. *PLoS Genet.* **12**, e1006293 (2016).

32. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

33. Fu, C. et al. Multiple aspects of male germ cell development and interactions with Sertoli cells require inositol hexakisphosphate kinase-1. *Sci. Rep.* **8**, 7039 (2018).

34. Inoue, H. et al. Mouse D1Pas1, a DEAD-box RNA helicase, is required for the completion of first meiotic prophase in male germ cells. *Biochem. Biophys. Res. Commun.* **478**, 592–598 (2016).

35. Barral, S. et al. Histone variant H2A.L.2 guides transition protein-dependent protamine assembly in male germ cells. *Mol. Cell* **66**, 89–101 e8 (2017).

36. Yamada, D. et al. Disruption of spermatogenic cell adhesion and male infertility in mice lacking TSLC1/IGSF4, an immunoglobulin superfamily cell adhesion molecule. *Mol. Cell Biol.* **26**, 3610–3624 (2006).

37. Wakayama, T. & Iseki, S. Role of the spermatogenic-Sertoli cell interaction through cell adhesion molecule-1 (CADM1) in spermatogenesis. *Anat. Sci. Int.* **84**, 112–121 (2009).

38. Ozaki-Kuroda, K. et al. Nectin couples cell-cell adhesion and the actin scaffold at heterotypic testicular junctions. *Curr. Biol.* **12**, 1145–1150 (2002).

39. Bronson, R., Mikhailik, A., Schwedes, J., Gnatenko, D. & Hatchwell, E. Detection of candidate nectin gene mutations in infertile men with severe teratospermia. *J. Assist. Reprod. Genet.* **34**, 1295–1302 (2017).

40. Cortez, D. et al. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493 (2014).

41. Chen, Y. et al. Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res.* **28**, 879–896 (2018).

42. Braun, R. E., Behringer, R. R., Peschon, J. J., Brinster, R. L. & Palmiter, R. D. Genetically haploid spermatids are phenotypically diploid. *Nature* **337**, 373–376 (1989).

43. Ernst, C., Eling, N., Martinez-Jimenez, C. P., Marioni, J. C. & Odom, D. T. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat. Commun.* **10**, 1251 (2019).

44. Bhutani, K. et al. Widespread haploid-biased gene expression enables sperm-level natural selection. *Science* **371**, eabb1723 (2021).

45. Daish, T. J., Casey, A. E. & Grutzner, F. Lack of sex chromosome specific meiotic silencing in platypus reveals origin of MSCI in therian mammals. *BMC Biol.* **13**, 106 (2015).

46. Zhou, Y. et al. Platypus and echidna genomes reveal mammalian biology and evolution. *Nature* **592**, 756–762 (2021).

47. Mahadevaiah, S. K. et al. Key features of the X inactivation process are conserved between marsupials and eutherians. *Curr. Biol.* **19**, 1478–1484 (2009).

48. Sin, H. S. & Namekawa, S. H. The great escape: active genes on inactive sex chromosomes and their evolutionary implications. *Epigenetics* **8**, 887–892 (2013).

49. Guioli, S., Lovell-Badge, R. & Turner, J. M. Error-prone ZW pairing and no evidence for meiotic sex chromosome inactivation in the chicken germ line. *PLoS Genet.* **8**, e1002560 (2012).

50. VanInsberghe, M., van den Berg, J., Andersson-Rolf, A., Clevers, H. & van Oudenaarden, A. Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. *Nature* **597**, 561–565 (2021).

51. Attar, M. et al. A practical solution for preserving single cells for RNA sequencing. *Sci. Rep.* **8**, 2151 (2018).

52. Krishnaswami, S. R. et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* **11**, 499–524 (2016).

53. Yates, A. et al. Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).

54. Marin, R. et al. Convergent origination of a *Drosophila*-like dosage compensation mechanism in a reptile lineage. *Genome Res.* **27**, 1974–1987 (2017).

55. Cunningham, F. et al. Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751 (2019).

56. Yates, A. D. et al. Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2019).

57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* **17**, 10–12 (2011).

58. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).

59. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

60. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

61. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).

62. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).

63. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 (2019).

64. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).

65. Green, C. D. et al. A comprehensive roadmap of murine spermatogenesis defined by single-cell RNA-seq data resources. *Dev. Cell* **46**, 651–667 (2018).

66. Lukassen, S., Bosch, E., Ekici, A. B. & Winterpacht, A. Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Sci. Rep.* **8**, 6521 (2018).

67. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

68. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).

69. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

70. Cassa, C. A. et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).

71. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

72. Van Der Lee, R., Wiel, L., Van Dam, T. J. P. & Huynen, M. A. Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. *Nucleic Acids Res.* **45**, 10634–10648 (2017).

73. Kosiol, C. et al. Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**, e1000144 (2008).

74. Shao, Y. et al. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome Res.* **29**, 682–696 (2019).

75. Futschik, M. E. & Carlisle, B. Noise-robust soft clustering of gene expression time-course data. *J. Bioinform. Comput. Biol.* **3**, 965–988 (2005).

76. Hooper, S. E., Weller, H. & Amelon, S. K. Countcolors, an R package for quantification of the fluorescence emitted by *Pseudogymnoascus destructans* lesions on the wing membranes of hibernating bats. *J. Wildl. Dis.* **56**, 759–767 (2020).

77. Weller, H. I. & Westneat, M. W. Quantitative color profiling of digital images with earth mover's distance using the R package colordistance. *PeerJ* **7**, e6398 (2019).

78. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

79. Dimitrov, D. et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat. Commun.* **13**, 3224 (2022).

80. Flaquer, A., Rappold, G. A., Wienker, T. F. & Fischer, C. The human pseudoautosomal regions: a review for genetic epidemiologists. *Eur. J. Hum. Genet.* **16**, 771–779 (2008).

81. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).

82. Babcock, M., Yatsenko, S., Stankiewicz, P., Lupski, J. R. & Morrow, B. E. AT-rich repeats associated with chromosome 22q11.2 rearrangement disorders shape human genome architecture on Yq12. *Genome Res.* **17**, 451–460 (2007).

83. R Core Team. *R: A Language and Environment for Statistical Computing*, https://www.R-project.org/ (R Foundation for Statistical Computing, 2018).

84. Bakken, T. E. et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).

Acknowledgements

We thank all members of the Kaessmann group, S. Tirier for discussions and N. Trost for the administration of the Kaessmann laboratory server. Computations were performed on the Kaessmann laboratory server and the bwForCluster from the Heidelberg University Computational Center (supported by the state of Baden-Württemberg through bwHPC and the German Research Foundation grant no. INST 35/1134-1 FUGG). This research was supported by grants from the ERC (grant no. 615253, OntoTransEvol) and German Research Council (DFG, grant nos. SFB 873 and KA 1710/4-1) to H.K., by the CellNetworks Postdoc Fellowship and EMBO Long-Term Fellowship to F.M. (grant no. ALTF 591-2017), and by the Australian Research Council (grant no. FT160100267) to F.G. and by

the Novo Nordisk Foundation (grant no. NNF21OC0069913) to K.A. and (grant no. NNF18OC0031004) to M.H.S. The use of all other mammalian samples for the type of work described in this study was approved by ERC ethics screening panels (ERC starting grant no. 242597, SexGenTransEvolution and ERC consolidator grant no. 615253, OntoTransEvol).

Author information

Author notes

1. These authors contributed equally: Florent Murat, Noe Mbengue

Authors and Affiliations

1. Center for Molecular Biology (ZMBH), DKFZ-ZMBH Alliance, Heidelberg University, Heidelberg, Germany

   Florent Murat, Noe Mbengue, Evgeny Leushkin, Mari Sepp, Julia Schmidt, Celine Schneider, Katharina Mößinger, Thoomke Brüning, Francesco Lamanna & Henrik Kaessmann

2. INRAE, LPGP, Rennes, France

   Florent Murat

3. Department of Growth and Reproduction, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark

   Sofia Boeg Winge & Kristian Almstrup

4. International Center for Research and Research Training in Endocrine Disruption of Male Reproduction and Child Health, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark

   Sofia Boeg Winge & Kristian Almstrup

5. Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark

Sofia Boeg Winge, Meritxell Riera Belles & Mikkel Heide Schierup

6. Berlin Institute of Health at Charité, University of Medicine Berlin, Corporate Member of the Free University of Berlin, Humboldt-University of Berlin, Berlin, Germany

   Timo Trefzer & Christian Conrad

7. Evolutionary Developmental Biology Laboratory, Francis Crick Institute, London, UK

   Margarida Cardoso-Moreira

8. Biomedical Primate Research Center (BPRC), Rijswijk, the Netherlands

   Ivanela Kondova & Ronald Bontrop

9. German Primate Center (DPZ), Platform Degenerative Diseases, Göttingen, Germany

   Rüdiger Behr

10. German Center for Cardiovascular Research (DZHK), Partner Site Göttingen, Göttingen, Germany

    Rüdiger Behr

11. Center for Neurobiology and Brain Restoration, Skolkovo Institute of Science and Technology, Moscow, Russia

    Philipp Khaitovich

12. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

    Svante Pääbo

13. Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Spain

Tomas Marques-Bonet

14. Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain

Tomas Marques-Bonet

15. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

Tomas Marques-Bonet

16. Miquel Crusafont Catalan Institute of Paleontology, Autonomous University of Barcelona, Barcelona, Spain

Tomas Marques-Bonet

17. The Robinson Research Institute, School of Biological Science, University of Adelaide, Adelaide, South Australia, Australia

Frank Grützner

18. Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Kristian Almstrup

Contributions

F.M., N.M. and H.K. conceived and organized the study based on H.K.'s original design. F.M., N.M. and H.K. wrote the manuscript, with input from all authors. F.M. performed all analyses, developed the shiny app and drew the species icons. K.M. and T.B. collected the samples. R. Behr, P.K, F.G, S.P., I.K., R. Bontrop, T.M.-B., K.A. and M.H.S. provided samples. F.M., N.M., T.T. and M.S. established snRNA-seq methods. N.M. performed the snRNA-seq experiments with support from J.S. C.S., K.M. and T.B. S.B.W. performed the smISH experiments, and analysed the results together with

K.A. E.L. performed the testis genomic annotations. M.S., M.C.-M, F.L., M.R.B., C.C., K.A. and M.H.S. provided useful feedback and discussions.

Corresponding authors

Correspondence to Florent Murat, Noe Mbengue or Henrik Kaessmann.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Camille Berthelot, Sue Hammoud and Susana Chuva de Sousa Lopes for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Cellular composition and marker gene assessment.

a–e, Clustering and UMAP representations (left), marker gene-based cell type assignment (right), for **a**, primates, **b**, mouse, **c**, opossum, **d**, platypus and **e**, chicken. *CLU* marks Sertoli cells; *TAGLN* and *ACTA2* peritubular

and smooth muscle cells; *CD34* and *TM4SF1* endothelial cells; *APOE* and *CD74* macrophages; *STAR* and *CYP11A1* Leydig cells; *GFRA1, PIWIL4* (undifferentiated), *DMRT1*(differentiated), and *STRA8* spermatogonia; *SYCE1* (leptotene), *SYCP1* (zygotene), *PIWIL1* (pachytene)*, SYCP2, TANK* and *AURKA* spermatocytes; *LRRIQ1* (early), *ACRV1* (late) and *SPACA1* (late) round spermatids; *SPATA3*, *NRBP1, PRM1* and *GABBR2* elongated spermatids. **f**, Cell-type composition. **g**, Scaled expression of specific genes (y-axis) for all species and cell types (x-axis).

Extended Data Fig. 2 Mammalian testicular gene expression divergence.

a, Gene expression phylogenies. Neighbor-joining trees based on pairwise expression level distances (1−rho, Spearman's correlation coefficient) for the main testicular cell types, respectively ("Other somatic" refers to peritubular, smooth muscle, endothelial, macrophage, and Leydig cells as detailed in the legend of Extended Data Fig. 1 and Methods). Trees are drawn to the same scale (indicated by the scale bar). Bootstrap values (i.e., proportions of replicate trees that have the branching pattern as in the majority-rule consensus tree shown) are indicated by circles at the corresponding nodes: ≥ 0.9 (white fill). **b**, Human-macaque (left, $n = 304$) and human-mouse (right, $n = 257$) pairwise Spearman correlations after downsampling to the same number of nuclei and UMI across cell types for each species. Boxplots show the median (central value); upper and lower quartile (box limits) and 95% confidence intervals (whiskers) for 100 bootstrap replicates.

Extended Data Fig. 3 Evolutionary forces.

a, Tolerance to functional variants. **b**, Average normalized ratio of nonsynonymous over synonymous nucleotide substitutions (dN/dS) of expressed genes across primates. **c**, Percentage of positively selected genes along spermatogenesis. Opossum, platypus, and chicken were not used in the original studies[72,73] and so the 1:1 orthologues in these species were used. **d**, Phylogenetic age of cellular transcriptomes. Higher values reflect increased lineage-specific gene contributions (that is, younger transcriptomes). **e**, Percentages of UMIs that map to protein coding genes

(top) or to intergenic elements (bottom). XLOC_030429 drives a different mouse pattern. **f**, Translational efficiency (TE; normalized $\log_2$-transformed values) of expressed genes. **g**, Tissue specificity. **h**, Time specificity (development). In **a**, **b**, and **d**, human middle temporal gyrus (MTG) snRNA-seq data[84] was used for non-gonadal (brain) comparisons. See legend of main Fig. 2 for further details regarding the plotted measures. Box plots depict the median (center value); upper and lower quartile (box limits) with whiskers at 1.5 times the interquartile range.

Extended Data Fig. 4 Examples of gene expression trajectory changes along primate spermatogenesis validated by smISH.

Left: Median gene expression trajectory from snRNA-seq along spermatogenesis for human, bonobo, chimp., and gorilla (from top to bottom). For each replicate, mean expression levels across cells of a given cell type were calculated. Marks indicate values for the replicates with the highest and lowest mean expression levels, dots indicate the median of mean expression values of the replicates. There are two biological replicates for human, bonobo, and gorilla and three biological replicates for chimpanzee. Center: images of seminiferous tubule cross sections from human, chimp., and orangutan (from top to bottom) stained with smISH for *MYO3B* (**a**) and *ADAMTS17* (**b**) using RNAScope. Red asterisks indicate on which branch a trajectory change was called. Examples of spermatogonia (SG) and round spermatids (rSD) closeups are provided for visualization purposes. Right: percentage of positive cells in SG and rSD from 10 tubules per section (n = 3 for human, n = 1 for chimp. and orangutan). Box plots represent the mean (central value) distribution of staining; upper and lower quartile (box limits) with whiskers at 1.5 times the interquartile range.

Extended Data Fig. 5 Conservation of gene expression trajectories along spermatogenesis.

a, Global trajectory conservation score across amniotes for changed (n = 237) and conserved (n = 389) trajectories ($P < 2.2\text{x}10^{-16}$, two-sided Wilcoxon rank-sum test). **b**, Trajectory conservation score between human

and mouse of genes that lead to either a fertile (n = 1077) or an infertile (n = 108) phenotype when knocked out in mouse (*P* = 0.007, two-sided Wilcoxon rank-sum test). **a,b**, Box plots depict the median (center value); upper and lower quartile (box limits) with whiskers at 1.5 times the interquartile range. **c**, Gene expression trajectories (*DMRT1*, top; *IP6K1*, bottom) along primate spermatogenesis (spermatogenesis, from left to right). For each replicate, mean expression levels across cells of a given cell type were calculated. Marks indicate values for the replicates with the highest and lowest mean expression levels, dots indicate the median of mean expression values of the replicates. There are three replicates for chimpanzee; two replicates for human, bonobo, gorilla, macaque, and marmoset; one replicate for gibbon. The dashed vertical lines separate spermatogonia (SG), spermatocytes (SC), round (rSD) and elongated spermatids (eSD).

Extended Data Fig. 6 Functions of conserved and changed trajectories.

a, Top 20 enriched biological process GO terms of genes showing conserved expression trajectories across primates and peak expression in the different cell types, respectively. **b**, Top 20 enriched biological process GO terms of genes showing conserved trajectories between human and mouse and peak expression in the different cell types, respectively. **c**, The top 20 enriched biological process GO terms of genes with trajectory changes in primates. **d**, The top 20 enriched biological process GO terms in genes showing trajectory changes in amniotes. **e**, Human tissue (left) and time (right) specificity for changed (n = 237) and conserved (n = 389) trajectories across amniotes, respectively (two-sided Wilcoxon rank-sum tests were performed for statistical comparisons). Box plots depict the median (center value); upper and lower quartile (box limits) with whiskers at 1.5 times the interquartile range.

Extended Data Fig. 7 New genes with crucial spermatogenic functions.

a, percentages of genes among new genes of different ages (i.e., emergence in last common vertebrate, tetrapod, mammalian, eutherian, or rodent ancestors) that show infertility phenotypes when knocked out in the mouse (numbers of genes with infertility phenotypes and total number of genes considered for an age category are provided above each bar). **b**, Expression level trajectories of the retrogenes *H2al2a* and *D1Pas1* in spermatogenic and somatic cell types. For each of the two replicates, mean expression levels across cells of a given cell type were calculated. Marks indicate values for the replicates and dots indicate the median of the values of the replicates. **c**, Trees of *H2al2a* and *D1Pas1* and paralogs from which they originated through the mechanism of RNA-based gene duplication (i.e., *D1Pas1* stems from *Ddx3y*, whereas the precise ancestral parental gene of *H2al2a* is not easily discernable due to a complex history of a number of RNA- and DNA-based duplication events).

Extended Data Fig. 8 Sertoli-germ cell communications mediated by ligand-receptor interactions in testis across mammals.

a, Significant ligand-receptor interactions (as assessed by CellPhoneDB[78]; see Methods) across species for Sertoli-spermatogonia, Sertoli-spermatocytes, Sertoli-round spermatids and Sertoli-elongated Spermatids communications (details in Supplementary Table 9). **b**, Overview of all distinct significant ligand-receptor interactions between Sertoli and the four principal germ cell types across species. The number of detected significant interactions for each species in both panels is indicated to the bottom left of each plot (Set size).

Extended Data Fig. 9 Per chromosome testis-specific genes enriched in spermatogonia and Sertoli cells.

a,**c**, Per chromosome percentage of expressed testis-specific genes (y-axis) with predominant expression in differentiated spermatogonia (human and macaque) or spermatogonia (mouse, opossum and platypus) (**a**) and Sertoli cells (**c**) versus the percentage of all genes in the genome located on a given chromosome (x-axis: “% genes genome-wide”; i.e., the number of genes on

a given chromosome divided by the total number of genes in the genome). X chromosomes are colored in red. The diagonal shows the numbers expected if testis-specific genes were randomly distributed across the genome. **b**, Location of platypus testis-specific genes enriched in spermatogonia on chromosomes X1, X3 and X5.

Extended Data Fig. 10 Classification of X and Y bearing spermatids.

The scatter and bar plots show the percentages of X and Y transcripts, and the distribution of X transcripts (%), respectively, across nuclei or cells in our single-nucleus (left) and publicly available single-cell data in mouse[43] (**a**) and human[23] (**b**). The red and green lines depict the fitted curves for the bimodal distributions.

Extended Data Fig. 11 Transcriptome differences between X and Y bearing spermatids.

The heatmaps indicate transcript abundances of significantly enriched genes in X- and Y-bearing spermatids in the different species (a: human, b: mouse, c: opossum, d: platypus). The corresponding bar plots (genes in the same order) show the genes with significantly enriched expression in X- and Y-bearing cells, respectively using two-sided Wilcoxon rank-sum tests with Bonferroni correction (ordered according to increasing *P*-values) (see also lists in Supplementary Table 11). Genes marked with an asterisk are long non-coding RNAs. Genes in red are located on autosomes, those in black on sex chromosomes. “XLOC” denotes loci annotated in our previous work[6]; notably, only few XLOC loci (opossum: XLOC-076551, XLOC-076500, XLOC-076152, XLOC-076546) were found to have coding potential (as assessed by analyses based on ribosome profiling data generated in our previous work[6]), whereas all others likely represent lncRNAs.

Extended Data Fig. 12 Mammalian sex chromosome evolution.

a, Y-to-autosome transcript ratios in individual germ cells along spermatogenesis. For spermatids, only Y-bearing cells are considered.

Whiskers depict the median ± 25th and 75th percentiles. **b**, Average expression along platypus X chromosomes in spermatogonia and pachytene spermatocytes. Genes are grouped into 1,000,000 bp bins along the chromosomes. **c**, To identify potential MSCI escapees, we screened for X-linked genes in SDRs with a significant increase in transcript abundance from SG to SC stages subject to MSCI, which ensures that potential escape genes are indeed actively transcribed in SC (i.e., they do not merely represent genes expressed in SG with stable transcripts still detectable in SC) (Methods). Expression differences ($\log_e$-fold change, x-axis) of X and autosomal genes between SC and SG were assessed using differential expression analysis (two-sided Wilcoxon rank-sum tests; *p*-values on the Y-axis are Bonferroni-corrected). Vertical red lines indicate a 0.25-fold expression change ($\log_e$-scale); horizontal red lines indicate corrected $P < 0.05$. Potential escapees (i.e., X-linked genes in SDRs with significantly higher expression in SC than SG) are indicated by red arrows.

Supplementary information

Supplementary Information

Supplementary Figs. 1–5, which show technical aspects of QC and ISH quantifications, and the legends for Supplementary Tables 1–11 (in a separate Excel file).

Reporting Summary

Peer Review File.

Supplementary Tables

Supplementary Tables 1–11; see main PDF for legends.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Murat, F., Mbengue, N., Winge, S.B. *et al.* The molecular evolution of spermatogenesis across mammals. *Nature* **613**, 308–316 (2023). https://doi.org/10.1038/s41586-022-05547-7

- Received: 02 November 2021
- Accepted: 09 November 2022
- Published: 21 December 2022
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05547-7

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

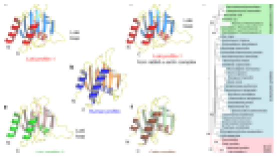
Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

Exploring the rapid evolution of the mammalian testis

Nature Research Briefing 21 Dec 2022

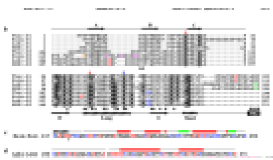
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05547-7



Genomes of Asgard archaea encode profilins that regulate actin

03 October 2018

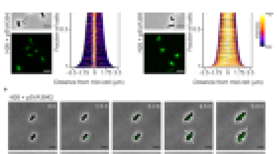
Caner Akıl & Robert C. Robinson



Structural and biochemical evidence for the emergence of a calcium-regulated actin cytoskeleton prior to eukaryogenesis

31 August 2022

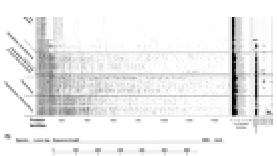
Caner Akıl, Linh T. Tran, … Robert C. Robinson



The archaeal protein SepF is essential for cell division in Haloferax volcanii

08 June 2021

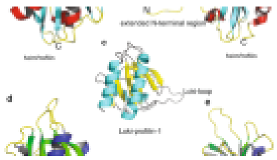
Phillip Nußbaum, Maren Gerstner, … Sonja-Verena Albers



[Identification and characterization of novel filament-forming proteins in cyanobacteria](#)

05 February 2020

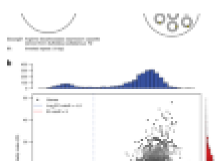
Benjamin L. Springstein, Christian Woehle, … Karina Stucken



[Heimdallarchaea encodes profilin with eukaryotic-like actin regulation and polyproline binding](#)

01 September 2021

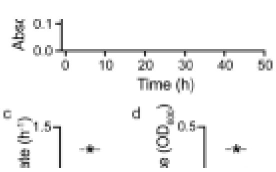
Sabeen Survery, Fredrik Hurtig, … Celestine N. Chi



[The essential genome of the crenarchaeal model Sulfolobus islandicus](#)

21 November 2018

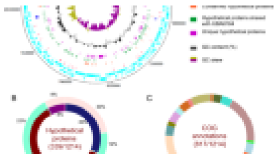
Changyi Zhang, Alex P. R. Phillips, … Rachel J. Whitaker



[Nutrient complexity triggers transitions between solitary and colonial growth in bacterial populations](#)

17 March 2021

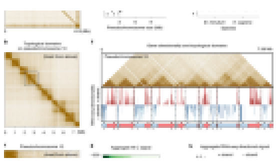
Glen G. D’Souza, Vanessa R. Povolo, … Martin Ackermann



[Bodo saltans (Kinetoplastida) is dependent on a novel Paracaedibacter-like endosymbiont that possesses multiple putative toxin-antitoxin systems]

15 January 2021

Samriti Midha, Daniel J. Rigden, … Andrew P. Jackson



[Transcription-dependent domain-scale three-dimensional genome organization in the dinoflagellate Breviolum minutum]

29 April 2021

Georgi K. Marinov, Alexandro E. Trevino, … William J. Greenleaf

- Article
- Open Access
- Published: 21 December 2022

Actin cytoskeleton and complex cell architecture in an Asgard archaeon

- Thiago Rodrigues-Oliveira,
- Florian Wollweber,
- Rafael I. Ponce-Toledo,
- Jingwei Xu,
- Simon K.-M. R. Rittmann,
- Andreas Klingl,
- Martin Pilhofer &
- …
- Christa Schleper

Show authors

Nature **volume 613**, pages 332–339 (2023)

- 30k Accesses
- 686 Altmetric
- Metrics details

Abstract

Asgard archaea are considered to be the closest known relatives of eukaryotes. Their genomes contain hundreds of eukaryotic signature proteins (ESPs), which inspired hypotheses on the evolution of the eukaryotic cell[1,2,3]. A role of ESPs in the formation of an elaborate cytoskeleton and complex cellular structures has been postulated[4,5,6], but never visualized. Here we describe a highly enriched culture of '*Candidatus* Lokiarchaeum ossiferum', a member of the Asgard phylum, which thrives anaerobically at 20 °C on organic carbon sources. It divides every 7–14 days, reaches cell densities of up to 5×10^7 cells per ml and has a significantly larger genome compared with the single previously cultivated Asgard strain[7]. ESPs represent 5% of its protein-coding genes, including four actin homologues. We imaged the enrichment culture using cryo-electron tomography, identifying '*Ca*. L. ossiferum' cells on the basis of characteristic expansion segments of their ribosomes. Cells exhibited coccoid cell bodies and a network of branched protrusions with frequent

constrictions. The cell envelope consists of a single membrane and complex surface structures. A long-range cytoskeleton extends throughout the cell bodies, protrusions and constrictions. The twisted double-stranded architecture of the filaments is consistent with F-actin. Immunostaining indicates that the filaments comprise Lokiactin—one of the most highly conserved ESPs in Asgard archaea. We propose that a complex actin-based cytoskeleton predated the emergence of the first eukaryotes and was a crucial feature in the evolution of the Asgard phylum by scaffolding elaborate cellular structures.

Main

Soon after the discovery of archaea as a separate lineage besides bacteria, molecular and phylogenetic studies suggested that there is a deep common evolutionary descent between archaea and eukaryotes[8,9]. However, only recently has the discovery of the first Lokiarchaeota[10] (now Lokiarchaeia[11]) and the wider superphylum of Asgardarchaeota in metagenomic analyses[1,2,11,12,13,14,15,16,17] corroborated a distinct relationship and a possible direct emergence of eukaryotic cells from archaea. In fact, eukaryotes form a direct sister group to Asgardarchaeota or even arise within Asgardarchaeota in most phylogenomic analyses[3,10,18].

Compellingly, all members of the Asgardarchaea carry an extensive repertoire of genes that were originally assumed to be unique to eukaryotes (ESPs)[1,2,3,10,19,20]. These ESPs are mostly associated with features of cells with high complexity, such as cytoskeleton formation, transport and the shaping of membranes. For example, the observation that Asgard archaeal genomes encode a complete and functional ubiquitin-coupled ESCRT system[10,21] suggested the possibility of elaborate intracellular membrane compartments[10]. Another notable example is genes encoding several close homologues of eukaryotic actin. While F-actin-like assemblies have been identified in other archaea[22,23], Asgard archaea also possess actin-related proteins (ARPs), as well as actin-binding proteins. Notably, Asgard profilins and gelsolins were found to be able to modulate the dynamics of eukaryotic actin[5,19,24,25,26], indicating the existence of an elaborate and dynamic

cytoskeleton[4]. However, the in situ structures and functions of archaeal actins remain unclear.

A seminal study[7] presented the first enrichment culture of an Asgard archaeon, '*Candidatus* Prometheoarchaeum syntrophicum', which grows slowly to low cell densities in syntrophic consortia with molecular-hydrogen-consuming organisms. As '*Ca.* P. syntrophicum' cells show long branched protrusions, the authors proposed a hypothesis for eukaryogenesis, in which a primordial Asgard archaeon closely interacts with the predecessor of the bacterial endosymbiont and eventually endogenizes it[7,27]. These observations were consistent with the stepwise mechanism of eukaryogenesis that was first proposed as the 'inside out' model[28]. The role of Asgard ESPs could so far not be investigated in the natural host, making it difficult to further test these conceptual models. Although '*Ca.* P. syntrophicum' was shown to transcribe ESPs, characteristics regarding its intracellular architecture could not be revealed. Fundamental questions regarding the presence of a cytoskeleton or internal compartmentalization in Asgard archaea remain unclear, as does the structure of the cell envelope. Here we combine the enrichment of an experimentally tractable Asgard archaeon with state-of-the-art imaging to reveal its cellular architecture at macromolecular detail.

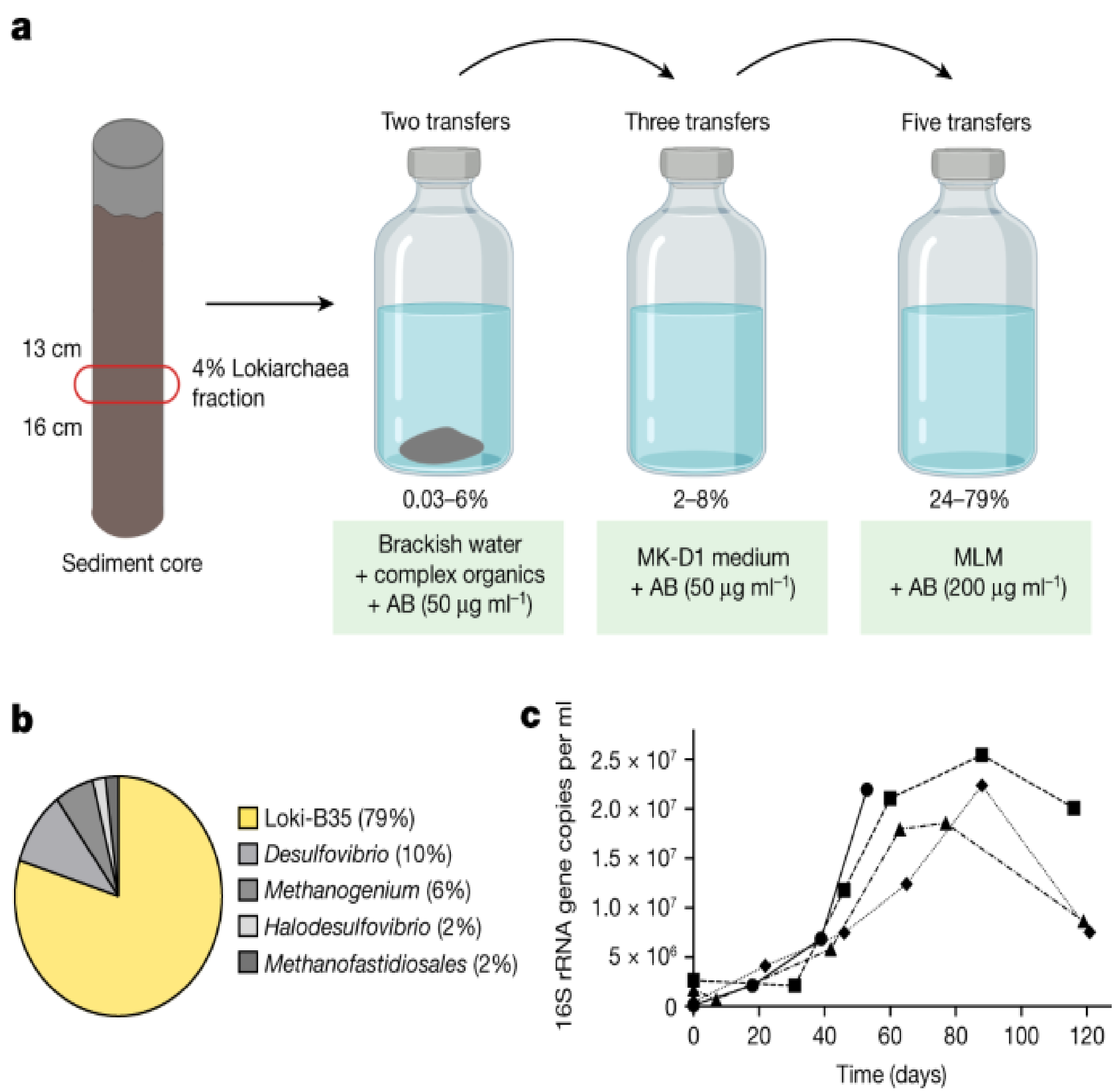
A Lokiarchaea culture from sediment

Considering that lokiarchaeal organisms and other Asgard archaea can be found in a variety of anoxic and often marine environments[29,30], we screened DNA from shallow-water sediment from different locations for the presence of 16S rRNA genes of Asgard archaea to select suitable and easily reachable sampling sites for establishing enrichments. Sediments from a small estuarine canal that regularly receives water from the Mediterranean near the coast of Piran, Slovenia, were identified to have the highest relative abundance at the 13–16 cm depth layer, exhibiting up to 4% of Asgard archaea 16S rRNA genes in amplicon sequencing (Extended Data Figs. 1 and 2).

The identified sample was used to inoculate enrichment cultures (Fig. 1a) with media of different compositions and various headspace conditions

(Supplementary Table 1). With periodic monitoring using quantitative PCR (qPCR) with Lokiarchaea-specific primers (Extended Data Fig. 3), growth could be observed after 140 days at 20 °C in serum flasks containing sterile-filtered water from the original source supplemented with complex organics (casein hydrolysate, milk powder and amino acids). However, after two transfers under these conditions, growth could no longer be detected and a second round of screening with different medium compositions was performed. Using a modification of the medium MK-D1 reported for the cultivation of '*Ca.* P. syntrophicum'[7], cell growth recovered, and abundances reached repeatedly 2–8%. However, higher enrichments, were not achieved under these conditions. Only through developing a minimal medium, mostly by reducing the input of organic carbon sources to a single compound and by increasing antibiotic concentrations, lokiarchaeal relative abundances reached between 25% and 80% after several transfers. The highest enrichments were achieved in minimized lokiarchaeal medium (MLM) with casein hydrolysate (Fig. 1a), while growth was also observed with either tryptone, peptone, milk powder, single amino acids, glucose or pyruvate (Supplementary Table 1).

Fig. 1: Enrichment and cultures of Loki-B35.



a, Schematic of our cultivation approach. A sediment core fraction was used as an inoculum for cultivation in sterile-filtered environmental water from the sampling site supplemented with complex organic compounds. The enrichment was then transferred to modified MK-D1 medium[7]. Enrichments of up to 79% were obtained when the cultures were transferred to MLM supplemented with casein hydrolysate. AB, antibiotics. The figure was created using BioRender. **b**, The composition of the culture with the highest enrichment as assessed by 16S rRNA gene amplicon sequencing. **c**, Growth curves ($n = 4$) of Loki-B35 in MLM (80:20 N_2:CO_2) supplemented with casein hydrolysate (growth was quantified by qPCR), indicating maximum cell densities of 2.5×10^7 per ml and generation times of about 7–14 days.

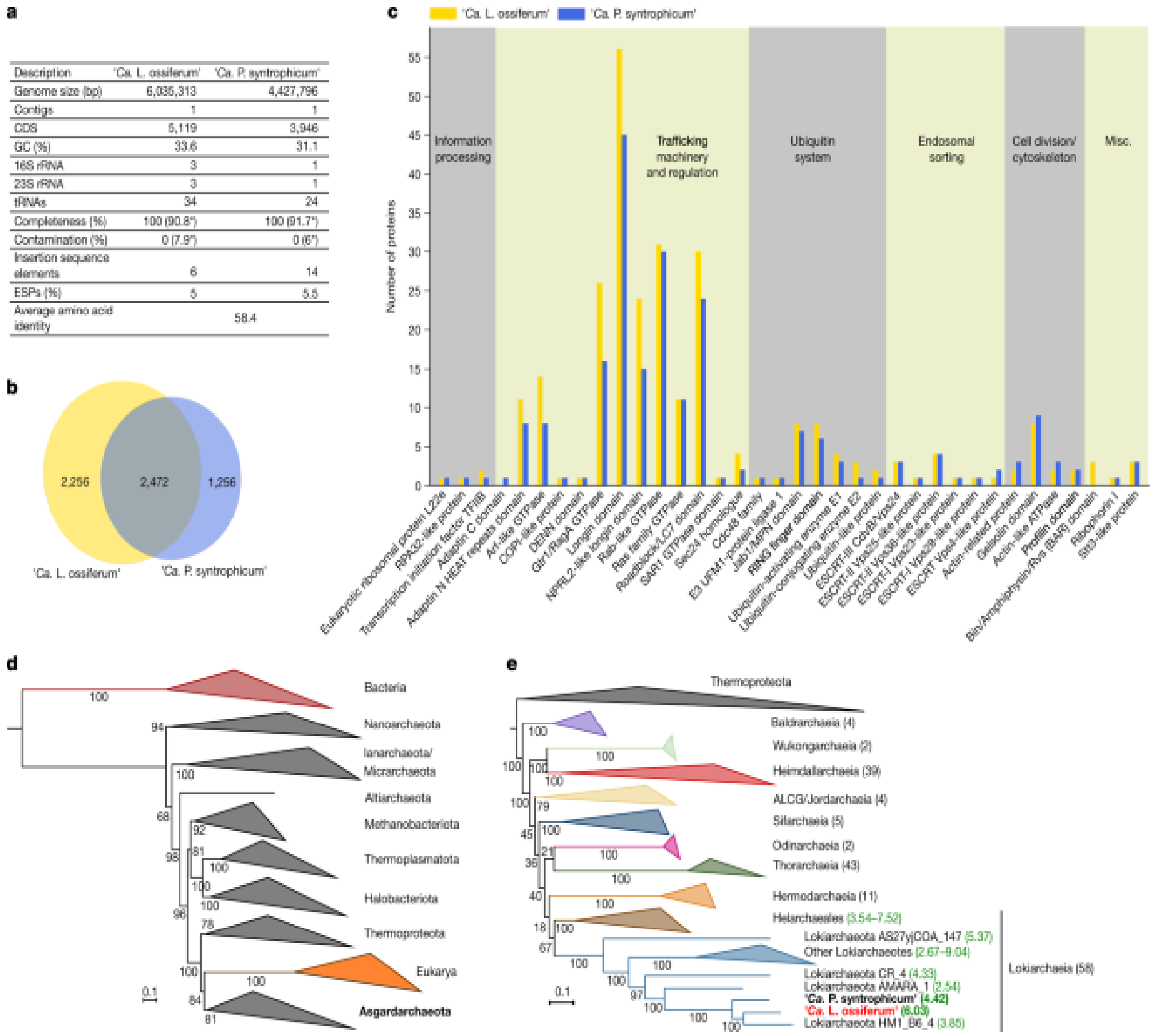
Amplicon sequencing analyses of 16S rRNA genes revealed that the culture with the highest enrichment (Loki-B35) consisted of three dominant and two minor species: a single Lokiarchaeon sequence (79%), a sulfate-reducing bacterium of the *Desulfovibrio* genus (10%), a hydrogenotrophic methanogen of the *Methanogenium* lineage (6%), as well as a *Halodesulfovibrio* and a member of the *Methanofastidiosales* genus (both at around 2%) (Fig. 1b and Extended Data Fig. 1). Notably, both *Halodesulfovibrio* and *Methanogenium* were also syntrophic partners in the enrichment cultures of '*Ca.* P. syntrophicum'[7], which stems from a geographically different and deep-sea environment. Thus, it seems probable that both Lokiarchaea rely on a similar metabolism, involving the fermentation of peptides to H_2 and/or small organic acids.

Loki-B35 grows without lag phase to maximum cell densities of up to 2.5×10^7 cells per ml, sometimes even 5×10^7 cells per ml within 50 to 60 days (Fig. 1c) when started with a 10% inoculum. However, extremely long lag phases of 90 to 120 days were observed when the inoculum originated from stationary cultures. The generation time was estimated to be 7 to 14 days. Compared with the deep sea Lokiarchaeon '*Ca.* P. syntrophicum' (maximum cell densities of 10^5 cells per ml, generation time of 14–25 days)[7], Loki-B35 grows considerably faster and to higher cell densities.

Loki-B35 genome analyses and phylogeny

The genome of Loki-B35 was assembled into one contig based on short and long read sequencing. It contains 6,035,313 base pairs (bp), encoding 5,119 predicted proteins, three 16S and 23S ribosomal rRNA copies (two of them in operons) and 34 tRNAs (Fig. 2a). The presence of only one rRNA operon in the closed genomes of '*Ca.* P. syntrophicum' and two Heimdallarchaea[7,31], but two operons in the high-quality assembly of the Lokiarchaeota genome '*Candidatus* Harpocratesius repetitus' FW102 (ref. [31]) and even three ribosomal RNA operons in Loki-B35, indicates that there is variability that may be based on the strains' generation time or flexibility to adapt to changing environmental conditions[32,33].

Fig. 2: Genome analysis and phylogenetic placement of ‘*Ca.* L. ossiferum’.



On the basis of the analysis of the closed genome of enrichment Loki-B35, we propose a description of the species ‘*Ca.* L. ossiferum’. **a**, The characteristics of the genome of ‘*Ca.* L. ossiferum’ in comparison to ‘*Ca.* P. syntrophicum’. Note the substantial difference in genome size. The values indicated by asterisks are the estimated values of contamination and completeness on the basis of the identification of marker genes performed by CheckM[53] (Methods). **b**, The diagram shows to scale the number of shared clusters of orthologous proteins between ‘*Ca.* L. ossiferum’ and ‘*Ca.* P. syntrophicum’ as well as genome-specific clusters. A more detailed analysis is provided in Extended Data Fig. 4. **c**, A comparison of the occurrence of ESPs in ‘*Ca.* L. ossiferum’ and ‘*Ca.* P. syntrophicum’.

Annotation of ESPs was performed according to the asCOG database[2]; general functional categories (on top) were added similar to earlier assignments[1]. Note that '*Ca.* L. ossiferum' is enriched for ESPs of the following protein families associated with trafficking machineries: adaptin N heat repeats domain; Arf-like GTPase; Gtr1/RagA GTPase; longin domain; NPRL2-like longin domain. **d**,**e**, Maximum-likelihood (ML) phylogenies based on the concatenation of 23 universally conserved ribosomal proteins[2]. **d**, Tree of life showing Eukarya as a sister clade of Asgardarchaeota. **e**, '*Ca.* L. ossiferum' and '*Ca.* P. syntrophicum' belong to the same Lokiarchaeia class. Taxonomic assignments are based on the recently proposed classification[11]. Branch supports were calculated with 1,000 ultrafast bootstrap samples. The values in square brackets show the genome sizes of complete genomes (bold) and MAGs within Lokiarchaeia.

Compared with '*Ca.* P. syntrophicum', the genome of Loki-B35 is significantly larger (by approximately 1.6 Mb), which is also reflected by 2,256 unique proteins (Fig. 2b). Functional annotation of orthogroups revealed that genes of Loki-B35 are enriched in almost all categories, reflecting its overall larger genome size (Extended Data Fig. 4 and Supplementary table 2). Notably, the fraction of proteins representing ESPs scales approximately with genome size, representing 5.5% in '*Ca.* P. syntrophicum' (218 ESPs) and 5% in Loki-B35 (258 ESPs; calculated according to the most recent asCOG database[2]) (Supplementary table 3). Compared to '*Ca.* P. syntrophicum', the genome of Loki-B35 is particularly enriched for genes associated with membrane trafficking and protein transport (Fig. 2c).

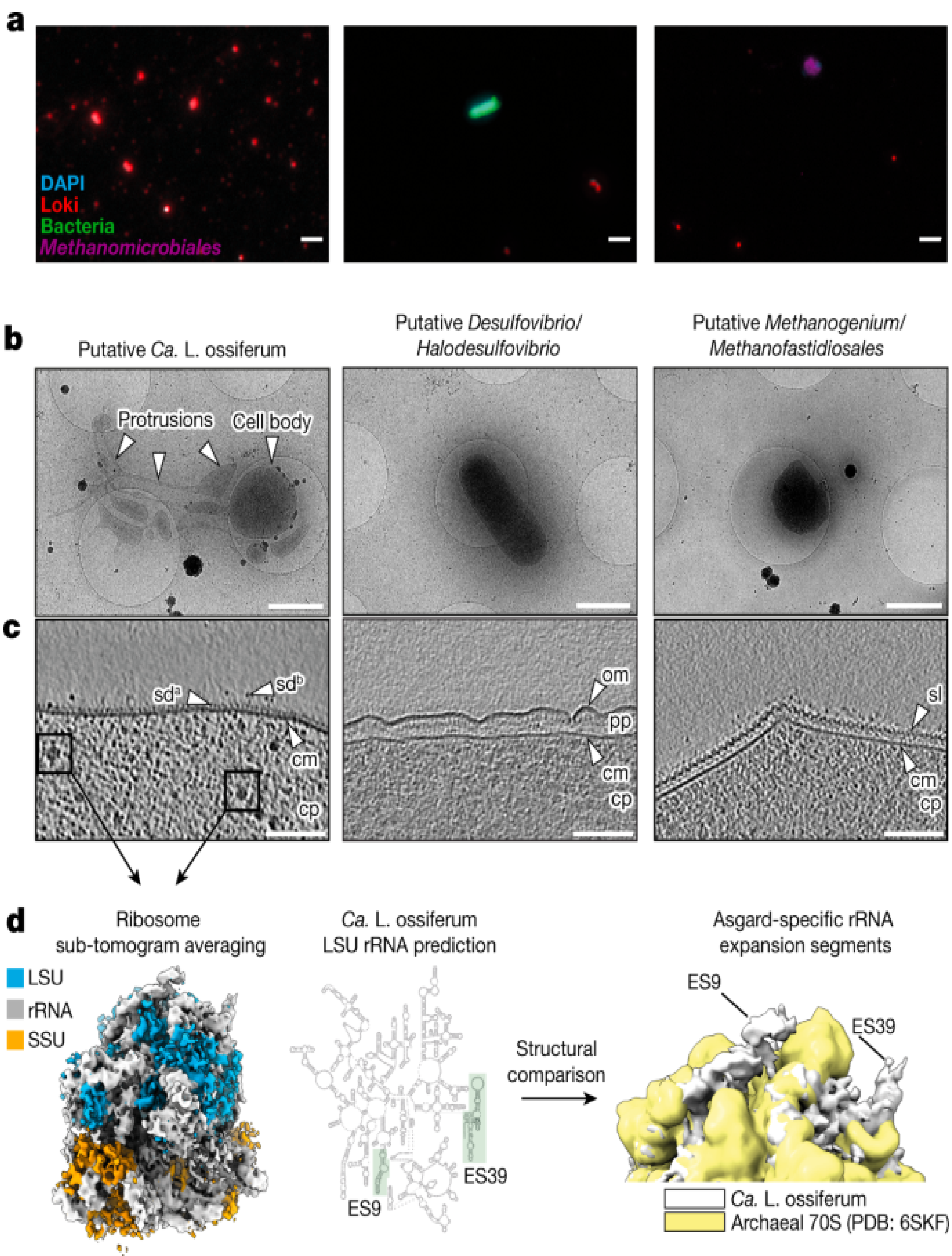
The sequence similarity of the 16S rRNA genes of Loki-B35 and '*Ca.* P. syntrophicum' is 95.3% and the common orthologous proteins share 58.4% amino acid identity, which justifies a separation into different genera[34,35]. This separation is supported by the large number of unique proteins that represent 47.7% of the complete predicted protein set. The evolutionary distance between the two cultivated Lokiarchaea also becomes evident in a gene synteny comparison, which shows a high number of rearranged genes and genome-specific regions being distributed throughout the genomes (Extended Data Fig. 5).

In a universal phylogeny based on 23 conserved ribosomal proteins from 291 representative species of all three domains of life, all Asgard archaea formed a monophyletic group with eukaryotes as their direct sister lineage (Fig. 2d). The Asgard phylogeny (Fig. 2e), based on 168 genomes, clearly separates all described classes of the phylum and is consistent with other recent phylogenomic analyses[1,2,13], although inner branching nodes vary depending on differences in the analyses and datasets. Loki-B35 together with '*Ca.* P. syntrophicum' and three other lokiarchaeal metagenome-assembled genomes (MAGs) formed one of two major sublineages within the class Lokiarchaeia (Fig. 2e). On the basis of these analyses, we propose a new genus and species for which we propose the name: '*Ca.* L. ossiferum' strain Loki-B35 (see below for etymology).

Identification of '*Ca.* L. ossiferum' cells

Our next goal was to identify individual '*Ca.* L. ossiferum' cells and characterize them using microscopy. Fluorescence in situ hybridization (FISH) with specific probes showed that '*Ca.* L. ossiferum' cells appeared as spheres of variable cell size (0.3–1.0 µm), being considerably smaller than other organisms in the enrichment (Fig. 3a). Cell counts from FISH analysis (not shown) confirmed the relative abundance of up to almost 80% in our highest enrichments as seen by amplicon sequencing (Fig. 1b).

Fig. 3: Identification of '*Ca.* L. ossiferum' cells in the enrichment culture.



a, Hybridization chain reaction-FISH analysis of the enrichment culture stained with DAPI (cyan) and nucleotide probes targeting the major species of the culture, that is, Lokiarchaea cells (red; the sample on the left was 70×

concentrated), bacteria (green) and *Methanomicrobiales* (purple). The FISH experiments were performed five independent times with similar results. Scale bars, 2 µm. **b**, Low-magnification 2D cryo-electron micrographs of the three major cell types that were observed after screening of the enrichment culture ($n = 2$ independent cultures), showing a putative '*Ca.* L. ossiferum' cell with a round cell body and complex cell protrusions (left), a Gram-negative bacterial cell (middle) and an archaeal cell (right). Scale bars, 1 µm. **c**, Slices through cryo-tomograms of all three organisms shown in **c** (slice thickness, 9.02 nm), detailing the characteristic cell envelope architecture of the three species. Putative Lokiarchaea show small and unordered surface densities (sd^a) and complex surface proteins (sd^b) protruding from a single membrane. cm, cytoplasmic membrane; cp, cytoplasm; om, outer membrane; pp, periplasm; sl, surface layer. Scale bars, 100 nm. **d**, Identification of '*Ca.* L. ossiferum' by Asgard-specific rRNA structures. Left, a sub-tomogram average (11.7 Å resolution) of ribosomes from cryo-tomograms of putative lokiarchaeal cells (large-subunit proteins (LSU), blue; small-subunit proteins (SSU), orange; rRNA, white). Middle, secondary structure prediction of the '*Ca.* L. ossiferum' large-subunit rRNA (expansion segments ES9/ES39 are labelled). Right, a superposition of the average with a low-pass filtered (11 Å) map of the *T. kodakarensis* 70S ribosome (Protein Data Bank (PDB): 6SKF; yellow). The '*Ca.* L. ossiferum' structure (white) shows prominent additional rRNA features that were identified as the Asgard-specific rRNA expansion segments ES9 and ES39. See also Extended Data Fig. 6.

We next plunge-froze cells of a live culture onto electron microscopy (EM) grids to image them in a near-native state using cryo-electron tomography (cryo-ET). A major challenge was the low cell density combined with the high fragility of the '*Ca.* L. ossiferum' cells, which did not allow us to perform any processing steps. Thus, instead of concentrating the samples, we performed an extensive screening of the grid by recording two-dimensional (2D) overview images, followed by cryo-ET data collection of selected cells. This approach revealed three general cell types that had distinct morphologies and cell envelope architectures (Fig. 3b,c). One class consisted of round-shaped cell bodies associated with elaborate and heterogeneous protrusions. The other two cell types were rods and spherically shaped cells, respectively, without protrusions. Their cell

envelopes had canonical Gram-negative and archaeal cell envelope features, therefore probably representing co-cultured bacteria and archaea (Fig. 3c). By contrast, the cell envelope of '*Ca.* L. ossiferum' candidates featured complex unordered densities protruding from a single membrane.

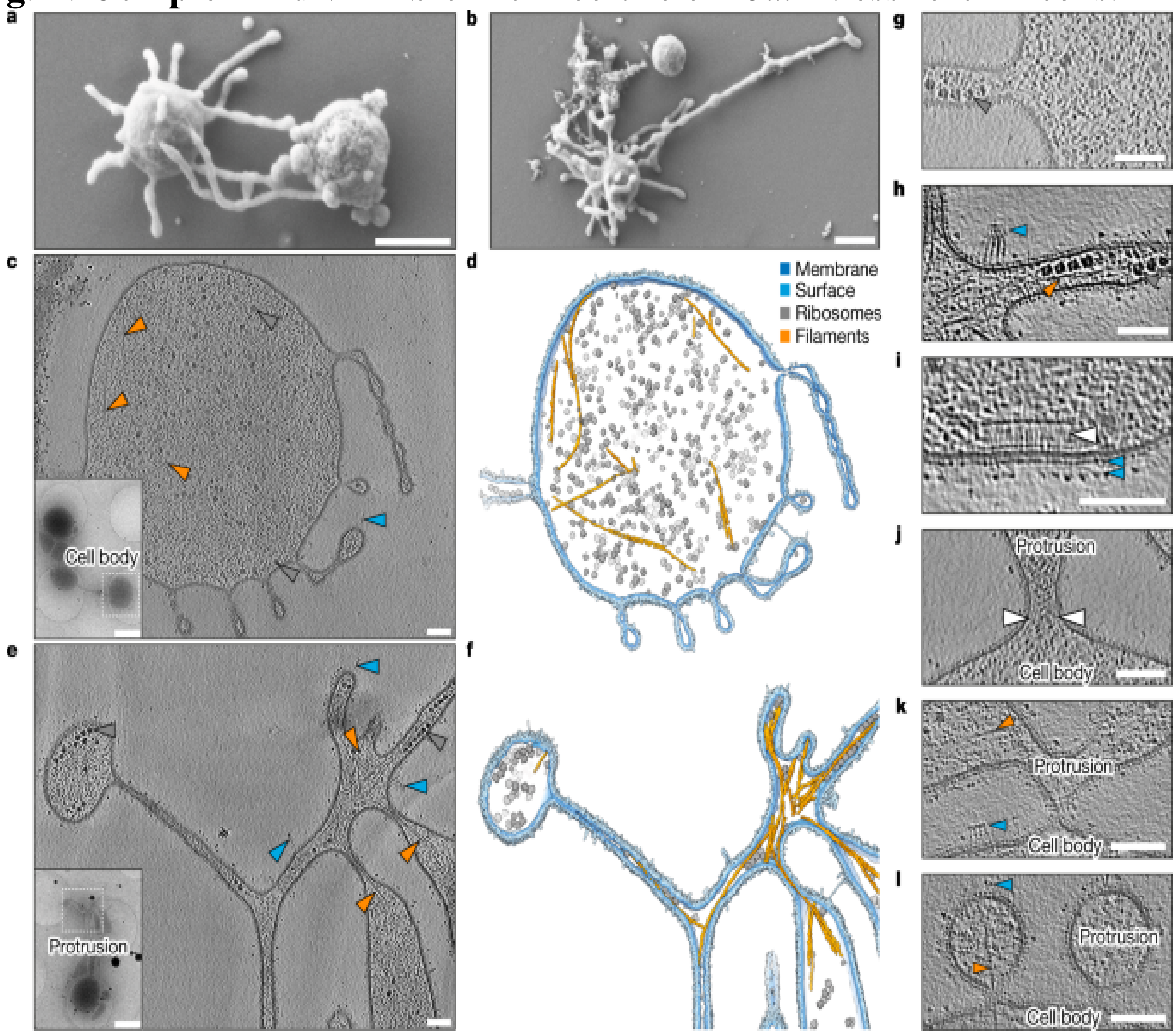
To unambiguously identify '*Ca.* L. ossiferum' cells, we initially attempted to correlate FISH with cryo-ET. FISH, however, involves harsh sample preparation steps (chemical fixation, dehydration, permeabilization and high temperatures), which did not allow for the preservation of the fragile cellular ultrastructure. We therefore turned to an alternative approach and hypothesized that '*Ca.* L. ossiferum' cells could be identified by unique structural features of their ribosomes. Supersized eukaryote-like expansion segments (ES9 and ES39) have been reported in Asgard archaeal rRNAs but are absent from bacteria and other archaea[36]. We therefore computationally extracted 4,126 sub-tomographic volumes of ribosomes from 56 tomograms of '*Ca.* L. ossiferum' candidate cell bodies and protrusions and averaged them (Fig. 3d). A comparison between the average and a high-resolution structure of a euryarchaeotal 70S ribosome revealed additional prominent densities, which we identified as Asgard-specific expansion segments ES9 and ES39, by correlating large subunit rRNA positions to the structural docking result (Fig. 3d and Extended Data Fig. 6c). Importantly, expansion segments were not present in the large-subunit rRNA sequences of co-cultured species (Extended Data Fig. 6d).

The cell type exhibiting a characteristic variable morphology, cell bodies with long protrusions and a single membrane with an elaborate surface proteome was therefore identified as '*Ca.* L. ossiferum'. These cells often appeared as individuals in FISH analyses (Extended Data Fig. 7a) and electron microscopy imaging (Fig. 3b). However, sometimes, '*Ca.* L. ossiferum' was found in aggregates with co-cultured species (Extended Data Fig. 7b,c). As these observations were rather infrequent, we assumed that, although nutrient exchange with co-cultured species may be necessary for the strain's growth, persistent cell–cell contact seems to not be obligate throughout its life cycle.

Complex and variable cell architecture

Having established the identity and general appearance of ‘*Ca.* L. ossiferum’ cells, we next aimed to analyse their overall organization by scanning EM. We identified small coccoid cells with surface-bound vesicles and extensive protrusions (Fig. 4a,b and Extended Data Fig. 7d). In contrast to ‘*Ca.* P. syntrophicum’, these long protrusions appeared more irregular, frequently branching or expanding into bulbous structures.

Fig. 4: Complex and variable architecture of ‘*Ca.* L. ossiferum’ cells.



a,**b**, SEM imaging of fixed ‘*Ca.* L. ossiferum’ cells showed small coccoid cells with extensive protrusions. Example micrographs from $n = 2$ independent cultures are shown. See also Extended Data Fig. 7d. For **a** and **b**, scale bars, 500 nm. **c**–**f**, Slices through cryo-tomograms (**c**,**e**; thickness, 9.02 nm) and the corresponding neural-network-aided 3D volume segmentations (**d**,**f**) of two different ‘*Ca.* L. ossiferum’ cells. The insets in **c**

and **e** show 2D overview images of the two different target cells. Cell bodies (**c**,**d**) and networks of protrusions (**e**,**f**) both contained ribosomes (grey arrowheads), cytoskeletal filaments (orange arrowheads) and complex surface densities (blue arrowheads). Note that **e** and **f** show the same cell as in Fig. 3c. For **c** and **e**, scale bars, 100 nm (tomogram) and 1 µm (2D overview). **g**,**h**, Expanded views of slices from tomograms in **c** and **e**, showing ribosome chains, complex surface proteins and filaments (colour code as in **c**–**f**) in a junction of a cell bridge (**g**) and a constricted part of the protrusion network (**h**). For **g** and **h**, scale bars, 100 nm. **i**–**l**, Slices through cryo-tomograms showing a putative chemoreceptor array (**i**; indicated by a white arrowhead) and different types of connections between cell bodies and protrusions (**j**–**l**). The coloured arrowheads indicate filaments and surface structures as defined for **c**–**f**. The white arrowheads in **j** indicate weak densities at the neck of the junction. Slice thickness, 9.02 nm (**j**) or 10.71 nm (**i** and **k**–**l**). For **i**–**l**, scale bars, 100 nm.

To investigate the macromolecular organization of '*Ca.* L. ossiferum' cells, we extended our cryo-ET analysis. In contrast to data obtained using scanning electron microscopy (SEM), membranous protrusions of unfixed cells showed even more elaborate shapes (Extended Data Fig. 7e). Some protrusions connected multiple larger cell bodies (Fig. 4c (inset)), reminiscent of cell bridges observed in other archaea[37,38,39,40]. Although the majority of cells exceeded the thickness limitation of cryo-ET, some were thin enough to visualize the internal architecture (Fig. 4c) in 3D, enabling us to not only observe large protein structures such as ribosomes, but also numerous thin and sometimes bent filaments (convolutional neural network-aided segmentations are shown in Fig. 4d,f; Supplementary Video 1). This network of filaments was even better resolved in some of the very thin (<100 nm) protrusions (Fig. 4e,f and Supplementary Video 1), which contained sometimes bundled filaments that connected different parts of the cell. Ribosomes were homogenously distributed throughout the cell body, cell bridges and protrusions (Fig. 4c–f), where they could sometimes be observed as membrane-associated chains (Fig. 4g). Except for two instances (Extended Data Fig. 7f), we did not observe internal membrane-bound compartments.

Our extended cryo-ET dataset also revealed further insights into the structure of the cell envelope. The single membrane was not only decorated

with a layer of small unordered densities, but also with a plethora of structures further protruding from the membrane (Fig. 4c–l (blue arrowheads)). Some densities connected different parts of the protrusion network (Fig. 4d,f), whereas others formed elaborate assemblies that localized to regions of high membrane curvature (Fig. 4f,h). On the cytoplasmic side of the membrane, we infrequently ($n = 2$) observed putative chemoreceptor arrays (Fig. 4i). Consistent with this observation, the genome of '*Ca.* L. ossiferum' encodes a set of chemotaxis proteins. Although absent from '*Ca.* P. syntrophicum', many of these are also present in other Lokiarchaeia and Heimdallarchaeia (in 27 out of 97 MAGs; Extended Data Fig. 8). The gene set includes 15 methyl-accepting chemotaxis proteins as well as *cheA/B/C/D/R/W/Y* (Supplementary Table 4). Together with the extensive repertoire of surface proteins, these may mediate cell–cell communication, interactions and motility.

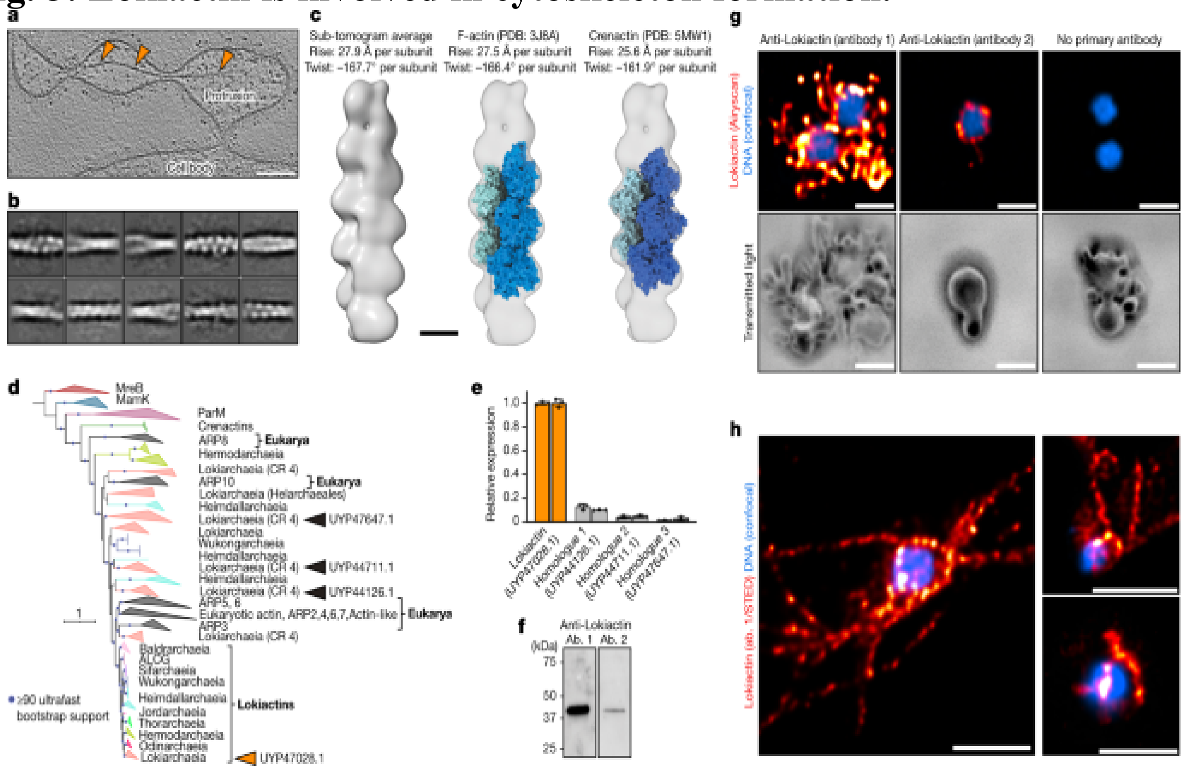
The unique cell envelope was mostly continuous between the cell body and protrusions, even though the transition zones showed high variability. Some appeared as stable junctions (Fig. 4j) (with potential densities appearing to stabilize the 'neck'), whereas others formed very thin constrictions (Fig. 4c,d) or were only loosely attached (Fig. 4k,l). Notably, single cytoskeletal filaments frequently traversed the junctions into the protrusions (Fig. 4l). In a similar manner, filaments also connected different parts of the protrusion network, often extending across constricted membrane tubes (Fig. 4f). These observations indicate that the cytoskeleton functions as a scaffold to maintain the elaborate cellular architecture of '*Ca.* L. ossiferum'.

Lokiactin-based cytoskeleton

To resolve the identity of the most frequently observed cytoskeletal filament in cryo-tomograms (Figs. 4 and 5a), we set out to determine its structure in situ and developed a workflow using helical reconstruction of 2D-projected sub-tomograms, followed by sub-tomogram averaging (Extended Data Fig. 9). 2D classification showed a two-stranded filament structure (Fig. 5b). The final reconstruction at a resolution of 24.5 Å enabled us to determine the helical parameters (rise of 27.9 Å per subunit, twist of −167.7 Å per subunit), which are highly similar to eukaryotic F-actin but also to the archaeal Crenactin[23]. Importantly, the dimensions of F-actin[41] and

Crenactin[23] filaments would be consistent with our reconstructed map (Fig. 5c).

Fig. 5: Lokiactin is involved in cytoskeleton formation.



a, Slice through a cryo-tomogram showing a cytoskeletal filament inside a protrusion at higher magnification. Slice thickness, 5.36 nm. Scale bar, 100 nm. **b**, Filament segments were extracted from cryo-tomograms for structural analysis. 2D classes that were obtained after helical reconstruction of 2D-projected filament particles are shown, indicating a twisted double-stranded architecture. Box size, 34.3 × 34.3 nm. See also Extended Data Fig. 9. **c**, Sub-tomogram average (24.5 Å resolution) of the cytoskeletal filament displaying helical parameters with a high similarity to eukaryotic F-actin and archaeal Crenactin. Structural docking shows that an F-actin-like filament is consistent with the reconstructed map. See also Extended Data Fig. 9. Scale bar, 50 Å. **d**, Maximum-likelihood phylogenetic tree of actin family proteins. The ‘*Ca.* L. ossiferum’ genome encodes four homologues. One homologue (GenBank: UYP47028.1) clusters together with other Asgard archaeal Lokiactins (group indicated by orange arrowhead), which form a sister group to eukaryotic actin. The three other homologues (from top to bottom: GenBank UYP47647.1, UYP44711.1 and UYP44126.1) cluster with other

Asgard archaeal and eukaryotic ARPs (groups indicated by the black arrowheads). The tree was rooted with the MreB protein family. CR 4, subgroup from within Lokiarchaeia. **e**,**f**, Lokiactin is expressed in the enrichment culture. **e**, Transcription of the four actin homologues was analysed using RT–qPCR analysis of two enrichment cultures, indicating the highest levels of transcripts for Lokiactin. The mean expression levels normalized to Lokiactin are shown. The error bars indicate the s.d. of three technical replicates. **f**, Expression was also detected using western blotting analysis of a cell lysate obtained from the enrichment culture (representative result from $n = 2$ independent samples). Gel source data are provided in Supplementary Fig. 1. Two antibodies (ab.) were used that were raised against different '*Ca.* L. ossiferum' Lokiactin-specific peptides. **g**,**h**, Immunofluorescence staining of '*Ca.* L. ossiferum' cells with two different Lokiactin-specific antibodies analysed using Airyscan (**g**) or stimulated emission depletion (STED) (**h**) imaging (representative images of $n = 3$ (**g**) or $n = 2$ (**h**) independent preparations). The distribution of fluorescent signal indicates the presence of Lokiactin-based cytoskeletal structures in cell bodies and protrusions, being consistent with observations from cryo-tomograms. The top row of **g** shows single slices of the fluorescent DNA signal (blue, Hoechst stain, LSM-Plus-processed confocal) and the Alexa Fluor 647-labelled secondary antibodies (red/orange, jDCV-processed Airyscan). The bottom row of **g** shows the minimum intensity *z*-projections of the transmitted light channel to indicate the cell shape. The control (right column) was probed with only secondary antibodies (the contrast in the top row was adjusted equally). The images in **h** show single slices of representative deconvolved STED images detecting Lokiactin (red/orange, abberior STAR 580-labelled secondary antibodies) and DNA (blue, SPY505-DNA). For **g** and **h**, scale bars, 1 μm.

The '*Ca.* L. ossiferum' genome contains four homologues of eukaryotic actin. One of these homologues clusters inside a group of Lokiactins, which form a sister group to bona fide eukaryotic actin and most ARPs in phylogenies (Fig. 5d). The remaining three homologues cluster with other ARPs from eukaryotes and Asgard archaea. Their phylogenetic patterns and distribution indicate a complex evolution within Asgard archaea, which seems to be shaped by gene losses, gains and duplications. By contrast, Lokiactin represents the most highly conserved group of actin homologues.

It is found in all Asgard lineages[5], implicating its presence in the last common ancestor of the entire phylum (Fig. 5d).

To test whether the actin-like cytoskeletal filament in '*Ca.* L. ossiferum' contained Lokiactin, we first tested the expression levels of all actin homologues using qPCR with reverse transcription (RT–qPCR) and found that Lokiactin expression was severalfold higher (Fig. 5e). We next generated two antibodies that were raised against Lokiactin-specific peptides. Using these antibodies, expression was detected in lokiarchaeal cell lysates by western blotting (Fig. 5f) and we observed staining throughout '*Ca.* L. ossiferum' cells in immunogold-labelling experiments (Extended Data Fig. 10). In immunofluorescence experiments, both antibodies revealed filamentous signals in the cell bodies and particularly also in protrusions, consistent with the abundant distribution of filaments observed in cryo-tomograms (Fig 5g,h). We therefore conclude that '*Ca.* L. ossiferum' possesses a complex Lokiactin-based cytoskeleton. As '*Ca.* L. ossiferum' encodes eight gelsolin-like and three profilin-like proteins that have been show to affect actin polymerization dynamics[4,5,19,24,25], we hypothesize that the assembly of the complex Lokiactin cytoskeleton in '*Ca.* L. ossiferum' is probably dynamically regulated by actin-binding proteins.

Conclusion

In conclusion, our comparatively fast-growing lokiarchaeal culture enabled us to study its cell architecture in a near-native state. We discovered an elaborate actin-based cytoskeleton in Asgard archaea, which has long been hypothesized[4,5,6,42,43,44], but has not been visualized.

The cytoskeleton is probably a hallmark structure of all Asgard archaea, as Lokiactin is conserved in genomes of all Asgardarchaeota classes. This clearly differentiates Lokiactin from many other ESPs that show complex evolutionary histories of gene gains and losses and lateral transfer resulting in patchy distributions in Asgard and also other archaea[2]. The high degree of Lokiactin conservation indicates strong constraints on its function. It is therefore probable that a dynamic cytoskeleton, regulated by numerous actin-related and actin-binding proteins, had a substantial impact on the

emergence, evolution and diversification of Asgard archaea. Our cryo-ET and immunofluorescence data revealed actin filaments in both parts of the cell. In the cell body, filaments are often found at the periphery, and they follow the longitudinal axis in the cell's protrusions. We therefore propose that Lokiactin acts as a scaffold for the complicated cell architecture of Asgard archaea, similar to eukaryotic actin, which is a major determinant of eukaryotic cell shape[45].

The elaborate cell architecture with extensive membranous protrusions has multiple implications for Asgard physiology and ecology. As these characteristic features make the cells highly fragile, it could also explain why the highest abundance of Asgard archaea is found in sediments[30,46,47,48] rather than in plankton. Importantly, our study established approaches that will enable imaging of Asgard archaea in a culture-independent manner in environmental samples, with the possibility of identifying Asgard cells in cryo-tomograms based on unique ribosomal RNA expansion segments.

The large surface area of the convoluted network of protrusions, in combination with the unusual cell envelope lacking a highly ordered S-layer (as typically found in other archaea) but rather displaying numerous surface proteins, may have enabled the intricate cell–cell contacts required for eukaryogenesis that—considering the lifestyle of the two cultured Asgard strains—probably involved interspecies dependencies in syntrophic relationships[49,50,51,52]. These findings strongly support a gradual path of mitochondrial acquisition through protrusion-mediated cell–cell interactions, which have been proposed previously in the inside-out and E3 hypotheses[7,28]. Additional experimental data—in particular, from diverse Asgard archaea—will be needed to further test these models and exclude alternative views[42,43]. Importantly, the cell architecture may enable the compartmentalization of cellular processes even in the absence of the internal organelle-like membrane systems that had been hypothesized based on the genomes of the first Asgard archaea[10].

Finally, our enrichment culture will serve as a model system to study the peculiar cell biology of Asgard archaea, as it grows to cell densities and purities that make it experimentally tractable. The intricate cell architecture of '*Ca.* L. ossiferum' suggests elaborate mechanisms for processes such as

macromolecular trafficking and assembly, membrane shaping, cell division and spatiotemporal regulation of the cytoskeleton, many of which are probably mediated by ESPs that can now be studied in situ.

Etymology. *Lokiarchaeum* (Lo.ki.ar.chae'um. N.L. neut. n. *archaeum* (from Gr. masc. adj. *archaios*, ancient), an archaeon; N.L. neut. n. *Lokiarchaeum*, an archaeon named after Loki, a god in Norse mythology). *ossiferum* (os.si'fe.rum. L. neut. pl. n. *ossa*, skeleton; L. v. *fero*, to carry; N.L. neut. adj. *ossiferum*, skeleton-carrying). The name describes a skeleton-carrying archaeon of the provisional class Lokiarchaeia[11] within the Asgardarchaeota phylum.

Locality. Isolated from a shallow sediment of an estuarine canal in Piran, Slovenia.

Diagnosis. Anaerobic archaeon of the Asgardarchaeota phylum that grows in enrichments with H_2-consuming bacteria and archaea at 20 °C on organic medium. Maximal cell densities reach 5×10^7 cells per ml at relative enrichments of up to 79%. Cells grow extensive, often branched protrusions with blebs and exhibit complex, irregular surface structures. The genome is 1,607,517 bp larger than that of the closest relative '*Ca.* P. syntrophicum' and shows 58.4% similarity at the amino acid level.

Methods

Sample collection and enrichment cultivation

Sediment core samples (length, approximately 60 cm) were retrieved from a shallow brackish canal (at 40 cm water depth, 19.5 °C, pH 8) near the coast of Piran, Slovenia (45° 29′ 46.1″ N 13° 36′ 10.0″ E) on 21 April 2019.

The sediment core was cut inside an anaerobic tent (N_2 atmosphere) at intervals of 3 cm, with each fraction being placed inside 50 ml conical sterile centrifuge tubes, sealed and stored at 4 °C. Around 0.5 g of sediment from the different fractions was used for DNA extraction and subsequent qPCR assays targeting lokiarchaeal 16S rRNA genes (qPCR conditions are provided below). The 13–16 cm deep sediment layer with the highest

relative 16S rRNA gene copies of Lokiarchaea per ml in relation to total DNA content (Extended Data Fig. 2) was selected for further use in cultivation, which was performed in 120 ml serum bottles sealed with butyl rubber stoppers. A total of 2 g of sediment was used as inoculum and different medium and headspace conditions were tested; growth was monitored using qPCR assays as described above. After 140 days, the growth was observed in cultures inoculated in 50 ml of sterile-filtered brackish water from the canal supplemented with casein hydrolysate (0.1% (w/v)), 20 amino acids (0.1 mM each) and milk powder (0.1%) incubated at 20 °C under an atmosphere of 80:20 N_2:CO_2 (0.3 bar). To try to limit bacterial growth, cultures were also supplemented with ampicillin, kanamycin and streptomycin (50 µg ml^{-1} each). The cultures were transferred into fresh medium whenever exponential growth could be detected (between 1–3 months). After two transfers, growth could no longer be observed under this set-up and the next transfer was performed in modified '*Ca.* P. syntrophicum' MK-D1 medium[7]. After three transfers under these conditions, the medium was reduced to further limit bacterial growth and, after five additional transfers in minimal medium, high enrichments were achieved. Eventually, the growth medium composition (per litre) was as follows: 20.7 g NaCl, 5 g $MgCl_2 \cdot 6H_2O$, 2.7 g $NaHCO_3$, 1.36 g $CaCl_2 \cdot 2H_2O$, 0.54 g NH_4Cl, 0.14 g KH_2PO_4, 0.03 g $Na_2S \cdot 9H_2O$, 0.03 g cysteine·HCl, 0.5 ml of acid trace element solution, 0.5 ml of alkaline trace element solution, 1 ml Se/W solution, 0.1% casein hydrolysate (w/v). The acid trace element solution contained (per litre): 1.491 g $FeCl_2 \cdot 4H_2O$, 0.062 g H_3BO_3, 0.068 g $ZnCl_2$, 0.017 g $CuCl_2 \cdot H_2O$, 0.099 g $MnCl_2 \cdot 4H_2O$, 0.119 g $CoCl_2 \cdot 6H_2O$, 0.024 g $NiCl_2 \cdot 6H_2O$ and 4.106 ml of HCl (37%). The alkaline trace element solution contained (per litre): 0.017 g Na_2SeO_3, 0.033 g Na_2WO_4, 0.021 g Na_2MoO_4 and 0.4 g NaOH. The medium pH was adjusted to 7.5 and contained ampicillin, kanamycin and streptomycin (200 µg ml^{-1} each). The headspace atmosphere was 80:20 N_2:CO_2 (0.3 bar) and the cultures were incubated at 20 °C without shaking.

DNA extraction and growth monitoring using qPCR

A total of 2 ml of the cultures was sampled every 14 days and centrifuged for 30 min at 20,000*g* at 4 °C. The supernatant was discarded and the resulting pellet was resuspended in 700 µl of SL1 buffer from the NucleoSpin Soil DNA extraction kit (Macherey-Nagel). The rest of the procedure was performed according to the manufacturer's instructions. High molecular mass DNA for genome sequencing was extracted using a standard phenol–chloroform-based protocol. The DNA concentration was measured with the Qubit 2.0 Fluorometer (Invitrogen), using the dsDNA HS kit, according to the manufacturer's instructions.

Lokiarchaea-specific 16S rRNA gene primers (LkF, 5′-ATCGATAGGGGCCGTGAGAG and LkR, 5′-CCCGACCACTTGAAGAGCTG) were designed using the ARB tool[54]. All assays were performed in triplicates on the CFX Connect Real-Time PCR Detection System (Bio-Rad) and data were collected using the CFX Maestro (v.2.3). Reaction mixtures (20 µl) contained: 1× Luna Universal qPCR Master Mix (New England BioLabs), 0.5 µM of each primer and 5–10 ng of template DNA. The cycling conditions were as follows: 95 °C for 1 min; then 40 cycles of 15 s at 95 °C and 1 min at 60 °C (for both annealing and extension), with a fluorescence reading after each cycle. Melting curves were generated by increasing the temperature from 60 °C to 95 °C, at 0.5 °C increments for 5 s with fluorescence readings after each increment. Lokiarchaeal 16S rRNA gene fragments amplified from sediment DNA were used as quantification standards. For quantification, triplicates of standard tenfold dilutions ranging from 10 to 10^8 copies were used in every assay. The efficiencies of these reactions varied from 90% to 100%, with R^2 values of >0.99. Primer specificity was confirmed through amplicon sequencing (Illumina MiSeq) of environmental DNA (Extended Data Fig. 3).

RNA extraction, cDNA synthesis and ARP RT–qPCR

From cultures, 20 ml was centrifuged at 20,000*g* for 30 min, 4 °C. The pellet was then resuspended in 600 µl of the lysis/binding buffer from the mirVana miRNA isolation kit (Invitrogen). The rest of the procedure was performed according to the manufacturer's instructions. Potential leftover genomic DNA was removed by incubating the samples with TURBO DNase

(Invitrogen) at 37 °C for 1 h. Lokiarchaeal 16S rRNA PCR tests were then used to confirm that no DNA had remained in the sample. cDNA was produced using the ProtoScript II First Strand cDNA Synthesis Kit according to the manufacturer's instructions.

Primers targeting Lokiactin (GenBank: UYP47028.1, 83F, 5′-GCAGGAGAAGATCAGCCTCG; 337R, 5′-AACCGGATGTTCGCTTGGAT) and actin homologue (GenBank: UYP44126.1; 82F, 5′-TGGGGGAGAAAATGAGCCAC; 443R, 5′-GGCCCCACGAACAGGATAAT), GenBank UYP44711.1 (361F, 5′-CCCTCCCAGACATTGCACAA; 731R, 5′-TGCGGGATCGACAGAATCAG) and GenBank UYP47647.1 (305F, 5′-CAAGGCTGGATCCCTTCAGA; 591R, 5′-ATTGCGTGATATGGTGGCCT) were designed using the Geneious Prime 2021.2 software (https://www.geneious.com). A temperature gradient PCR using culture DNA was used to determine the optimal annealing temperature and the temperature used for all primer pairs was 60 °C. The qPCR procedure and cycling conditions were the same as those used for lokiarchaeal 16S rRNA gene amplification. Amplified fragments from culture DNA were used as standards. The efficiencies of these reactions varied from 90% to 100%, with R^2 values of >0.99. Primer specificity was evaluated by melting curve analysis.

16S rRNA gene amplicon sequencing

Amplicon sequencing was performed by amplifying the extracted DNA using the general prokaryotic 16S rRNA gene targeting primer pair 515f (5′-GTGCCAGCMGCCGCGGTAA) and 806r (5′-GGACTACHVGGGTWTCTAAT)[55], which was then barcoded and sequenced at the Vienna BioCenter Core Facilities (VBCF) using the Illumina Miseq (300 PE) platform. Raw reads were processed using cutadapt[56] to remove primer sequences followed by the sequence analyses using the QIIME2 pipeline[57]. In brief, the DADA2 algorithm was used to denoise the data as well as to remove low-quality reads and chimeras. Sequences with 100% sequence identity were clustered into amplicon sequence variants. Taxonomy of amplicon sequence variants was assigned

using the SILVA database (release 138) with the q2-feature-classifier plugin[58].

In situ DNA-hybridization chain reaction

Cells were fixed in growth medium with the addition of 2.5% formaldehyde for 2 h at room temperature. After this period, they were washed three times in 1× phosphate-buffered saline and then stored at −20 °C in a mixture of absolute ethanol and 1× phosphate-buffered saline (1:1). The oligonucleotide probes used in this study were the general bacteria-targeting EUB338 (ref. [59]), the Methanomicrobiales-targeting MG1200 (ref. [60]) and the Lokiarchaea-specific DSAG-Gr2-1142 (ref. [7]) (Supplementary Table 5). The DNA-hybridization chain reaction procedure was performed as described previously[61] and the samples were imaged on the Eclipse Ni-U epifluorescence microscope (Nikon) using Gryphax (v.1.1.8.153). Images were processed using ImageJ[62].

Illumina sequencing

For short-read sequencing, metagenomic DNA of three samples from different time points in our enrichment was shotgun sequenced using the NovaSeq 6000 (paired-end, 150 bp) platform at Novogene. Sequencing data were processed using Trimmomatic (v.0.36)[63] to remove Illumina adapters and low-quality reads (SLIDINGWINDOW:5:20). The trimmed reads were co-assembled using SPAdes (v.3.15.2)[64] with the *k*-mer length varying from 21 to 111 and the '--meta' option. Contigs longer than 1,000 bp were binned using Binsanity[65], MaxBin2 (ref. [66]), MetaBAT[67] and CONCOCT[68] followed by contig dereplication and binning optimization using the DAS tool[69]. Completeness and contamination of MAGs was evaluated using CheckM[53] and taxonomic affiliation was obtained using GTDB-Tk (v.1.5.0; classify_wf)[70]. A lokiarchaeal MAG (genome length, 6,008,683 bp; completeness, 90.9%; contamination, 7.9%) was identified.

Nanopore sequencing and base calling

Library construction for ONT sequencing was performed using the SQK-LSK109 ligation kit followed by the PromethION sequencing using FLO-PRO002 flowcells. Fast5 files were processed using the high-accuracy ONT basecaller Bonito (v.0.3.6; https://github.com/nanoporetech/bonito) with the dna_r9.4.1 pretrained model (bonito basecaller --fastq dna_r9.4.1). Long-read demultiplexing and adapter removal was performed using Porechop (v.0.2.4)[71]. Reads shorter than 1 kb were removed using NanoFilt (v.2.8.0; NanoFilt -l 1000)[72]. We performed an initial metagenomic assembly of long reads using Flye (v.2.8.3-b1695)[73] with the '--meta' option. On the basis of the taxonomic assignment of contigs using the GTDB-Tk (v.1.5.0) classify workflow (classify_wf)[70], we identified a Lokiarchaeal genome with a total length of 6,035,157 bp in a single non-circular contig. To obtain the complete Lokiarchaeal genome, the following approach was applied. Long reads mapping to the short-read-assembled MAG and the long-read-assembled non-circular contig were extracted using minimap2 (ref. [74]) with the '-ax map-ont' option. Mapped reads were converted to BAM format using samtools[75], and bedtools bamtofastq[76] was used to obtain the reads in FASTQ format. Duplicated sequences were removed using the SeqKit[77] tool. After long-read deduplication, the remaining reads were assembled using flye (--nano-raw)[78] yielding a circular genome. Polishing and validation of the Loki-B35 genome was performed using four rounds of Pilon[79] as part of the validation pipeline of metagenomic assemblies proposed previously[80].

Phylogenomic tree

We recovered a total of 167 good-quality Asgard archaea genomes (completeness over 80% and less than 10% contamination) from the NCBI and other public databases (Supplementary Table 6). To reconstruct the phylogenomic tree of life, we collected a total of 23 ribosomal protein markers (Supplementary Table 7) from a wide range of bacteria, non-Asgard archaea and eukaryotes (Supplementary Tables 8 and 9) from a previous study[2].

The identification and retrieval of the 23 ribosomal markers in our Asgard genome database was based on the proteome annotation performed using

Prokka (v.1.14.6)[81]. Two phylogenomic trees were reconstructed in this study: (1) a tree of life with ribosomal markers from bacteria, eukaryotes, non-Asgard archaea and a subset of 94 Asgard genomes (Supplementary Table 6); and (2) a second tree using sequences only from Thermoproteota and the 168 Asgardarchaeota genomes present in our database. Ribosomal markers were aligned using the L-INSi algorithm of MAFFT (v.7.427)[82] and trimmed using BMGE using the default parameters[83]. Both sets of ribosomal markers were concatenated independently resulting in two multiple sequence alignment comprising 2,327 and 3,285 amino acid sites for the tree of life and the Thermoproteota plus Asgardarchaeota datasets, respectively. Phylogenomic reconstructions were performed with IQ-TREE 2 (ref. [84]) under the LG + C20 + F + G model with 1,000 ultrafast bootstrap replicates using UFBoot2 (ref. [85]).

Genome analysis and creation of orthogroups between '*Ca.* L. ossiferum' and '*Ca.* P. syntrophicum'

Protein sequences of '*Ca.* L. ossiferum' and '*Ca.* P. syntrophicum' were predicted using Prokka (v.1.14.6)[81]. The identified proteins were annotated using HMM searches against the Pfam database (v.34.0)[86] and KEGG Orthology was assigned using the BlastKOALA online server[87].

Proteomes of '*Ca.* L. ossiferum' and '*Ca.* P. syntrophicum' were used as queries in BLASTp searches ($e = 1 \times 10^{-10}$) against the recently published collection of Asgard clusters of orthologous genes[2] to obtain the asCOG numbers and annotation. Best matches were selected according to the following criteria: asCOG annotation of best hit was assigned if the BLAST alignment covered more than 70% of the query and sequence identity was greater than 25%; when the best BLAST hits did not meet the 70% threshold, a lower query coverage of 30% and at least 25% sequence identity was used to annotate protein domains. The Asgard COG annotation was used to identify ESPs on the basis of the curated ESP set composed of 505 AsCOGs previously described[2]. Average amino acid identity between '*Ca.* L. ossiferum' and '*Ca.* P. syntrophicum' was calculated using CompareM (v.0.1.2) (with the 'aai_wf' parameter)

(https://github.com/dparks1134/CompareM). Insertion sequence elements were identified using ISEScan[88].

Clusters of orthologous proteins between these two proteomes were identified using OrthoMCL (v.1.4)[89] according to the Synima pipeline[90] based on the all versus all protein alignments performed using BLASTp searches. Synima was also used to identify chains of syntenic genes between '*Ca.* L. ossiferum' and '*Ca.* P. syntrophicum' based on the DAGchainer software[91]. The genome synteny plot was constructed with Rideogram[92].

Phylogeny of actin-family proteins

Actin-like proteins of Asgardarchaeota genomes were identified on the basis of BLASTp searches using Lokiactins[1] as queries against our curated set of Asgard genomes ($e = 1 \times 10^{-10}$, sequence identity ≥ 25%). To retrieve a list of eukaryotic actin sequences, proteins were downloaded from UniProt based on the accession IPR004000, which corresponds to the actin domain present in the actin family. Proteins were clustered using CD-HIT[93] (-c 0.99 -d 0 -g 1) to remove highly similar sequences. Lokiactins were used as queries using BLASTp and only sequences with more than 25% identity were retained. Eukaryotic actins and Asgard actin-like proteins were aligned together with bacterial actin homologues involved cell shape determination (MreB), magnetosome organization (MamK) and plasmid segregation (ParM) based on the structural information of homologues of the DASH database using MAFFT (v.7.490; --dash --localpair --originalseqonly) followed by a trimming step with trimAl (-gappyout)[94]. The phylogenetic tree was reconstructed using IQ-TREE 2.0 (ref. [82]) under the LG + R8 model selected by ModelFinder[95].

SEM, conventional TEM and immunogold localization

For SEM, enrichment cultures of '*Ca.* L. ossiferum' were prepared as described previously[96]. SEM was performed using the Zeiss Auriga field-emission scanning electron microscope (Zeiss), operated at 2 kV.

For conventional resin embedding of samples dedicated for transmission electron microscopy (TEM) or immunogold localization, the samples were

high-pressure frozen and freeze-substituted using the Leica HPM 100 and AFS2 systems, respectively (Leica). The freeze substitution medium consisted of ethanol containing 0.5% glutaraldehyde, 0.5% formaldehyde and 0.5% uranyl acetate. The freeze substitution program, the following embedding in Epon epoxy resin and immunogold localization with Lokiactin-specific primary antibodies (ab1, see below) were performed using the protocol described previously for *Phaeodactylum tricornutum*[97]. Immunogold controls for primary and secondary antibody specificity were performed using *Vibrio harveyi* and *Pyrococcus furiosus*. Immunogold particle statistics were estimated for the controls as well as '*Ca.* L. ossiferum' (Supplementary Table 10). Transmission electron microscopy was performed using either the Zeiss EM 912 (Zeiss) system, operated at 80 kV and equipped with a Tröndle 2k × 2k slow-scan CCD camera (TRS, Tröndle Restlichtverstärker Systeme) or a JEOL F200 (JEOL), operated at 200 kV and equipped with a XAROSA 20 megapixel CMOS camera (EMSIS).

Generation of Lokiactin-specific antibodies and western blotting

The Lokiactin antibodies were raised against Lokiactin-specific peptides (ab1: CTFYTDLRVDPSEHPV; ab2: CSKNGFAGEDQPRSVF) and validated through ELISA assays using the services of Eurogentec. Peptide antibodies were designed by Eurogentec.

For western blotting, an aliquot of the culture was centrifuged at 20,000*g* for 10 min at 4 °C, the pellet was washed once in base MLM medium without casein hydrolysate and lysed and denatured in SDS-loading buffer (Bio-Rad) containing 1% (v/v) beta-mercaptoethanol at 95 °C. The samples were run on 4–20% Tris-Glycine gradient gels (Bio-Rad), transferred to PVDF membranes, blocked with 5% milk powder in TBST and probed with primary and secondary (goat anti-rabbit HRP, Invitrogen, 31460) antibodies. Signals were detected by enhanced chemiluminescence (ECL).

Immunofluorescence imaging

An aliquot of the culture was immobilized on poly-l-lysine-coated coverslips and fixed with 4% formaldehyde under a nitrogen atmosphere. Coverslips were blocked and permeabilized with 3% (w/v) BSA and 0.1% (v/v) Triton X-100 and subsequently probed with primary (anti-Lokiactin, 1:100 or 1:500 diluted) and secondary antibodies (either donkey anti-rabbit AF647, Invitrogen A-31573, 1:500 diluted; or goat anti-rabbit abberior STAR 580, abberior ST580-1002, 1:200 diluted) and counterstained with 10 µg ml^{-1} Hoechst 33342 (Thermo Fisher Scientific, for Airyscan imaging) or SPY505-DNA (Spirochrome, for STED samples). Coverslips were mounted with Vectashield (Vector Laboratories, Airyscan samples) or Prolong Diamond (Thermo Fisher Scientific, STED samples). The samples were imaged using a Zeiss LSM900 with Airyscan 2 detector and a ×63/1.4 NA oil-immersion objective. *z*-stacks of target cells were recorded using one confocal imaging track detecting the Hoechst signal and transmitted light and a separate Airyscan track detecting the Alexa Fluor 647 signal. Confocal stacks were deconvolved using the Zeiss LSM Plus processing function and Airyscan images were processed with Zeiss joint deconvolution (jDCV, 15 iterations) in ZenBlue (v.3.5). Minimum-intensity projections of the transmitted light channel and extraction of single confocal/Airyscan slices were performed in Fiji[98]. STED images were acquired using a Leica SP8 STED equipped with a ×100/1.4 NA oil-immersion objective. DNA and Lokiactin signals were detected in stack-wise sequential imaging tracks. Images were deconvolved with Huygens Professional (v.22.04; Scientific Volume Imaging; http://svi.nl).

Cryo-ET sample preparation

Samples were removed from the culture under a nitrogen atmosphere and mixed with 10 nm BSA-coated gold beads at a 1:5 ratio. The sample was kept in a nitrogen atmosphere until plunge-freezing. Then, 3.5 µl of the sample was applied to glow-discharged copper EM grids (R2/1, Quantifoil), automatically blotted from the backside (using a Teflon sheet on one side)[99] for 5–7 s and plunged into liquid ethane/propane[100] using the Vitrobot Mark IV (Thermo Fisher Scientific)[101].

Cryo-ET data collection

Cryo-ET data were collected on a Titan Krios G4 (Thermo Fisher Scientific) system operating at 300 kV equipped with a BioContinuum imaging filter and a K3 direct electron detector (Gatan). Data acquisition was performed using SerialEM[102,103]. Owing to the low cell density of the non-concentrated sample, grids were first extensively screened using polygon montages at low magnification (×2,250). After identification of targets, tilt series were acquired using a dose-symmetric tilt scheme[104], covering an angular range of −60° to +60° and a total electron dose of 140–160 e^{-} $Å^{-2}$. Tilt series were either acquired at a pixel size of 4.51 Å at the specimen level using 2° angular increments between tilts and a target defocus of −8 µm or at higher magnification (pixel size of 2.68 Å) with 3° angular increments and a defocus ranging from −3 to −6 µm (for sub-tomogram averaging of the ribosome and reconstruction of the cytoskeletal filament). 2D projection images shown in the Article were recorded at a magnification of ×2,250 (pixel size of 39.05 Å) and a target defocus of −200 µm.

Tomogram reconstruction, data processing and segmentation

Tilt series were drift-corrected using alignframes in IMOD[105] and 4×-binned tomograms were reconstructed by weighted-back projection in IMOD. To enhance the contrast for visualization and particle picking, tomograms were CTF-deconvolved and filtered using isonet[106]. 2D projection images were lowpass-filtered using mtffilter in IMOD. Segmentations were generated in Dragonfly (Object Research Systems, 2022; www.theobjects.com/dragonfly) as described previously[107]. In brief, isonet-filtered tomograms were further processed by histogram equalization and an unsharp filter, and a 5-class U-Net (with 2.5D input of 5 slices) was trained on 5–6 tomogram slices to recognize background voxels, filaments, membranes, cell surface structures and ribosomes. All neural-network-aided segmentations were cleaned up in Dragonfly, exported as a binary tiff and converted to mrc using tif2mrc in IMOD. Segmentations were visualized in ChimeraX[108].

Sub-tomogram averaging of the ribosome

Sub-tomogram averaging of ribosomes was performed using RELION (v.4.0)[109]. The individual ribosome particles were manually picked using

Dynamo[110] from 56 tomograms reconstructed at a binning factor of 4. The coordinates of particles and raw tilt series were imported into RELION (v.4.0) to generate pseudo sub-tomograms at a binning factor of 4 (4,126 particles). The particles were processed for 3D classification with a ribosome reference (Electron Microscopy Data Bank: EMDB-13448) that was low-pass filtered to 60 Å. The particles from good classes (class I and III in Extended Data Fig. 6) were processed for 3D auto-refinement at a binning factor of 2 and 3D local classification. The particles in the best 3D class (class II) were used for the 3D reconstruction at the binning factor of 1 and the resolution was further improved after three iterations of Tomo CTF refinement and frame alignment. The final structure of the ribosome at a resolution of 11.7 Å was reconstructed from 1,673 particles (Extended Data Fig. 6).

Identification of '*Ca.* L. ossiferum' by rRNA expansion segments

The ribosome sub-tomogram average was compared to a high-resolution structure of a Euryarchaeota ribosome (PDB: 6SKF, *T. kodakarensis*[111]) using fitinmap in ChimeraX to identify unique rRNA features. For visualization, the high-resolution structure was lowpass-filtered to 11 Å resolution using molmap in ChimeraX. Large subunit rRNA secondary structures of '*Ca.* L. ossiferum' and *T. kodakarensis* were predicted using R2DT[112] (available at https://rnacentral.org/r2dt) to identify the position of Asgard-specific rRNA expansion segments[36] and positions were mapped to the structural docking result to define ES9 and ES39 in the sub-tomogram average.

In situ reconstruction of cytoskeletal filaments

Filaments were analysed using a similar strategy as described in previous reports[113,114] and summarized in Extended Data Fig. 9a. In brief, individual filaments were manually picked from tomograms that were reconstructed at a binning factor of 4 using the filamentWithTorsion model in Dynamo. The filament segmentation was performed with an intersegment distance of 32.13 Å, resulting in a total of 10,031 segments from CTF-corrected and dose-weighted tomograms at the binning factor of 2. These segmented

subvolumes were processed for the filament analysis (Extended Data Fig. 9): (1) each subvolume was first rotated to make the helical axis parallel to the *z* axis based on the orientation calculated from particle coordinates (or derived from Dynamo) and then further centred by alignment with a featureless cylinder-like reference. (2) Re-oriented subvolumes were further rotated 90° to orient particles parallel to the *xy* plane. (3) The central slices of segments along the *z* axis were extracted and projected to generate a 2D projection dataset. (4) 2D projection images were processed for RELION helical reconstruction analysis[115] with an actin reference (EMDB-11976) low-pass filtered to 60 Å. (5) The orientations of 2D projection images were mapped back into the corresponding subvolumes and the polarity of the filament was validated from the orientations of segments on the same filaments, while the c entres of segments were refined in 3D space by aligning against the 3D model from 2D projection images using relion_refine. The polarity voting results were not applied during the reconstruction, as the resolution was not sufficient for clear polarity determination. The segments with refined orientations were then processed for a second iteration of filament analysis, in which the reference in the RELION analysis was updated to the 3D model from 2D projection images.

After three iterations of filament analysis, a 2D classification without sampling was performed in the last iteration to select only particles in 2D classes with visible features. The particles were then processed for 3D helical refinement to optimize the helical parameters in the process. A total of 3,161 2D projections from subvolumes were used to determine a reconstructed map at a resolution of 31 Å with refined helical parameters of 28.56 Å (rise/subunit) and −168.50° (twist/subunit). To further improve the map quality, we next performed sub-tomogram averaging using RELION (v.4.0)[109] (Extended Data Fig. 9a). Individual filaments were manually picked from 56 tomograms in Dynamo and were segmented with an intersegment distance of 32.13 Å. The coordinates of 13,569 filament segments and original raw tilt series were imported into RELION (v.4.0) to generate pseudo sub-tomograms at a binning factor of 2. The particles were processed for 3D auto-refinement without imposing helical symmetry and were then applied to 3D classification without angular sampling. The particles in the good 3D classes (class I and IV) were selected for next round of 3D refinement, in which the helical parameters from the previously

obtained 2D projection model were applied and optimized during refinement. The final structure at a resolution of 24.5 Å was reconstructed from 12,585 particles imposed with the refined helical parameters (rise = 27.9 Å, twist = −167.7°) (Fig. 5c and Extended Data Fig. 9).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The '*Ca.* L. ossiferum' genome sequence (accession CP104013) was uploaded to GenBank, under BioProject ID PRJNA847409, BioSample accession SAMN28933922. Sub-tomogram averages (EMD-15987–EMD-15988), representative tomograms (EMD-15989–EMD-15993) and the corresponding tilt series (EMPIAR-11269) have been uploaded to the Electron Microscopy Data Bank and the Electron Microscopy Public Image Archive.

References

1. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

2. Liu, Y. et al. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).

3. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **15**, 711–723 (2017).

4. Stairs, C. W. & Ettema, T. J. G. The archaeal roots of the eukaryotic dynamic actin cytoskeleton. *Curr. Biol.* **30**, R521–R526 (2020).

5. Akıl, C. et al. Mythical origins of the actin cytoskeleton. *Curr. Opin. Cell Biol.* **68**, 55–63 (2021).

6. Nobs, S. J., MacLeod, F. I., Wong, H. L. & Burns, B. P. Eukarya the chimera: eukaryotes, a secondary innovation of the two domains of life? *Trends Microbiol.* **30**, 421–431 (2022).

7. Imachi, H. et al. Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).

8. Huet, J., Schnabel, R., Sentenac, A. & Zillig, W. Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. *EMBO J.* **2**, 1291–1294 (1983).

9. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miytata, T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl Acad. Sci. USA* **86**, 9355–9359 (1989).

10. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

11. Sun, J. et al. Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota lineages. *ISME Commun.* **1**, 30 (2021).

12. Seitz, K. W., Lazar, C. S., Hinrichs, K. U., Teske, A. P. & Baker, B. J. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* **10**, 1696–1705 (2016).

13. Seitz, K. W. et al. Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat. Commun.* **10**, 1822 (2019).

14. Cai, M. et al. Diverse Asgard archaea including the novel phylum Gerdarchaeota participate in organic matter degradation. *Sci. China Life Sci.* **63**, 886–897 (2020).

15. Farag, I. F., Zhao, R. & Biddle, J. F. "Sifarchaeota," a novel Asgard phylum from Costa Rican sediment capable of polysaccharide degradation and anaerobic methylotrophy. *Appl. Environ. Microbiol.* **87**, e02584-20 (2021).

16. Zhang, J. W. et al. Newly discovered Asgard archaea Hermodarchaeota potentially degrade alkanes and aromatics via alkyl/benzyl-succinate synthase and benzoyl-CoA pathway. *ISME J.* **15**, 1826–1843 (2021).

17. Xie, R. et al. Expanding Asgard members in the domain of Archaea sheds new light on the origin of eukaryotes. *Sci. China Life Sci.* **65**, 818–829 (2022).

18. Bulzu, P.-A. et al. Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nat. Microbiol.* **4**, 1129–1137 (2019).

19. Akıl, C. et al. Insights into the evolution of regulated actin dynamics via characterization of primitive gelsolin/cofilin proteins from Asgard archaea. *Proc. Natl Acad. Sci. USA* **117**, 19904–19913 (2020).

20. Lu, Z. et al. Coevolution of eukaryote-like Vps4 and ESCRT-III subunits in the Asgard archaea. *Ecol. Evol. Sci.* **11**, e00417-20 (2020).

21. Hatano, T. et al. Asgard archaea shed light on the evolutionary origins of the eukaryotic ubiquitin-ESCRT machinery. *Nat. Commun.* **13**, 3398 (2022).

22. Ettema, T. J. G., Lindås, A. C. & Bernander, R. An actin-based cytoskeleton in archaea. *Mol. Microbiol.* **80**, 1052–1061 (2011).

23. Izoré, T., Kureisaite-Ciziene, D., McLaughlin, S. H. & Löwe, J. Crenactin forms actin-like double helical filaments regulated by arcadin-2. *eLife* **5**, e21600 (2016).

24. Akıl, C. & Robinson, R. C. Genomes of Asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).

25. Survery, S. et al. Heimdallarchaea encodes profilin with eukaryotic-like actin regulation and polyproline binding. *Commun. Biol.* **4**, 1024 (2021).

26. Inturi, R., Lara, S., Derweesh, M. & Chi, C. N. Structural characterization of a Thorarchaeota profilin indicates eukaryotic-like

features but with an extended N-terminus. *Adv. Biol.* **6**, e2101323 (2022).

27. Schleper, C. & Sousa, F. L. Meet the relatives of our cellular ancestor. *Nature* **577**, 478–479 (2020).

28. Baum, D. A. & Baum, B. An inside-out origin for the eukaryotic cell. *BMC Biol.* **12**, 76 (2014).

29. Manoharan, L. et al. Metagenomes from coastal marine sediments give insights into the ecological role and cellular features of *Loki*- and *Thorarchaeota*. *mBio* **10**, e02039-19 (2019).

30. Cai, M. et al. Ecological features and global distribution of Asgard archaea. *Sci. Total Environ.* **758**, 143581 (2021).

31. Wu, F. et al. Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized Asgard archaea genomes. *Nat. Microbiol.* **7**, 200–212 (2022).

32. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* **66**, 1328–1333 (2000).

33. Roller, B. R. K., Stoddard, S. F. & Schmidt, T. M. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nat. Microbiol.* **1**, 16160 (2016).

34. Luo, C., Rodriguez-R, L. M. & Konstantinidis, K. T. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* **42**, e73 (2014).

35. Konstantinidis, K. T., Rosselló-Móra, R. & Amann, R. Uncultivated microbes in need of their own taxonomy. *ISME J.* **11**, 2399–2406 (2017).

36. Penev, P. I. et al. Supersized ribosomal RNA expansion segments in asgard archaea. *Genome Biol. Evol.* **12**, 1694–1710 (2020).

37. Rosenshine, I., Tchelet, R. & Mevarech, M. The mechanism of DNA transfer in the mating system of an archaebacterium. *Science* **245**, 1387–1389 (1989).

38. Nickell, S., Hegerl, R., Baumeister, W. & Rachel, R. Pyrodictium cannulae enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *J. Struct. Biol.* **141**, 34–42 (2003).

39. Sivabalasarma, S. et al. Analysis of Cell–Cell Bridges in *Haloferax volcanii* using electron cryo-tomography reveal a continuous cytoplasm and S-layer. *Front. Microbiol.* **11**, 612239 (2021).

40. Marguet, E. et al. Membrane vesicles, nanopods and/or nanotubes produced by hyperthermophilic archaea of the genus *Thermococcus*. *Biochem. Soc. Trans.* **41**, 436–442 (2013).

41. Von Der Ecken, J. et al. Structure of the F-actin-tropomyosin complex. *Nature* **519**, 114–117 (2015).

42. Yutin, N., Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. The origins of phagocytosis and eukaryogenesis. *Biol. Direct* **4**, 9 (2009).

43. Bernander, R., Lind, A. E. & Ettema, T. J. G. An archaeal origin for the actin cytoskeleton: implications for eukaryogenesis. *Commun. Integr. Biol.* **4**, 664–667 (2011).

44. Van den Ent, F., Amos, L. A. & Löwe, J. Prokaryotic origin of the actin cytoskeleton. *Nature* **413**, 39–44 (2001).

45. Pollard, T. D. & Cooper, J. A. Actin, a central player in cell shape and movement. *Science* **326**, 1208–1212 (2009).

46. Yin, X. et al. Subgroup level differences of physiological activities in marine Lokiarchaeota. *ISME J.* **15**, 848–861 (2021).

47. Jørgensen, S. L., Thorseth, I. H., Pedersen, R. B., Baumberger, T. & Schleper, C. Quantitative and phylogenetic study of the deep sea

archaeal group in sediments of the arctic mid-ocean spreading ridge. *Front. Microbiol.* **4**, 299 (2013).

48. Jorgensen, S. L. et al. Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proc. Natl Acad. Sci. USA* **109**, E2846–E2855 (2012).

49. Moreira, D. & López-García, P. Symbiosis between methanogenic archaea and δ-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J. Mol. Evol.* **47**, 517–530 (1998).

50. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).

51. López-García, P. & Moreira, D. Eukaryogenesis, a syntrophy affair. *Nat. Microbiol.* **4**, 1068–1070 (2019).

52. Spang, A. et al. Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* **4**, 1138–1148 (2019).

53. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

54. Ludwig, W. et al. ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).

55. Caporaso, J. G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA* **108**, 4516–4522 (2011).

56. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).

57. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

58. Bokulich, N. A. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).

59. Amann, R. I., Krumholz, L. & Stahl, D. A. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *J. Bacteriol.* **172**, 762–770 (1990).

60. Raskin, L., Stromley, J. M., Rittmann, B. E. & Stahl, D. A. Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. *Appl. Environ. Microbiol.* **60**, 1232–1240 (1994).

61. Yamaguchi, T. et al. In situ DNA-hybridization chain reaction (HCR): a facilitated in situ HCR system for the detection of environmental microorganisms. *Environ. Microbiol.* **17**, 2532–2541 (2015).

62. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

63. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

64. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes de novo assembler. *Curr. Protoc. Bioinform.* **70**, e102 (2020).

65. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Binsanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **5**, e3035 (2017).

66. Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

67. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex

microbial communities. *PeerJ* **3**, 31165 (2015).

68. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

69. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

70. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2020).

71. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* **3**, e000132 (2017).

72. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

73. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

74. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

75. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

76. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

77. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).

78. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546

(2019).

79. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).

80. Giguere, D. J. et al. Complete and validated genomes from a metagenome. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.08.032540 (2020).

81. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

82. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

83. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

84. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

85. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

86. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

87. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).

88. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, 3340–

3347 (2017).

89. Li, L., Stoeckert, C. J. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).

90. Farrer, R. A. Synima: a synteny imaging tool for annotated genome assemblies. *BMC Bioinform.* **18**, 507 (2017).

91. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643–3646 (2004).

92. Hao, Z. et al. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020).

93. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

94. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

95. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

96. Brameyer, S. et al. Outer membrane vesicles facilitate trafficking of the hydrophobic signaling molecule CAI-1 between *Vibrio harveyi* cells. *J. Bacteriol.* **200**, e00740-17 (2018).

97. Flechsler, J. et al. 2D and 3D immunogold localization on (epoxy) ultrathin sections with and without osmium tetroxide. *Microsc. Res. Tech.* **83**, 691–705 (2020).

98. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

99. Weiss, G. L., Medeiros, J. M. & Pilhofer, M. in *Bacterial Protein Secretion Systems* (eds Journet, L. & Cascales, E.) 353–375 (Humana Press, 2017).

100. Tivol, W. F., Briegel, A. & Jensen, G. J. An improved cryogen for plunge freezing. *Microsc. Microanal.* **14**, 375–379 (2008).

101. Iancu, C. V. et al. Electron cryotomography sample preparation using the Vitrobot. *Nat. Protoc.* **1**, 2813–2819 (2007).

102. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

103. Schorb, M., Haberbosch, I., Hagen, W. J. H., Schwab, Y. & Mastronarde, D. N. Software tools for automated transmission electron microscopy. *Nat. Methods* **16**, 471–477 (2019).

104. Hagen, W. J. H., Wan, W. & Briggs, J. A. G. Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. *J. Struct. Biol.* **197**, 191–198 (2017).

105. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**, 71–76 (1996).

106. Liu, Y.-T. et al. Isotropic reconstruction for electron tomography with deep learning. *Nat. Commun.* **13**, 6482 (2022).

107. Hylton, R. K., Heebner, J. E., Grillo, M. A. & Swulius, M. T. Cofilactin filaments regulate filopodial structure and dynamics in neuronal growth cones. *Nat. Commun.* **13**, 2439 (2022).

108. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

109. Zivanov, J. et al. A Bayesian approach to single-particle electron cryo-tomography in RELION-4.0. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.28.482229 (2022).

110. Castaño-Díez, D., Kudryashev, M., Arheit, M. & Stahlberg, H. Dynamo: a flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments. *J. Struct. Biol.* **178**, 139–151 (2012).

111. Sas-Chen, A. et al. Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. *Nature* **583**, 638–643 (2020).

112. Sweeney, B. A. et al. R2DT is a framework for predicting and visualising RNA secondary structure using templates. *Nat. Commun.* **12**, 3494 (2021).

113. Martins, B. et al. Unveiling the polarity of actin filaments by cryo-electron tomography. *Structure* **29**, 488–498 (2021).

114. Wang, Z. et al. The molecular basis for sarcomere organization in vertebrate skeletal muscle. *Cell* **184**, 2135–2150 (2021).

115. He, S. & Scheres, S. H. W. Helical reconstruction in RELION. *J. Struct. Biol.* **198**, 163–176 (2017).

Acknowledgements

We thank M. Mousavian, M. A. Mardini, K. Hager and A. Tramontano for technical assistance in the cultivation process and M. Melcher for assistance in the initial discovery of Lokiarchaea in shallow sediments; F. Sousa for advice on phylogenetic and evolutionary aspects; N. Kroll and J. Grünert for technical assistance in immuno-gold labelling and SEM; A. Oren for advice in naming the organism; J. Heebner and M. Swulius (Penn State College of Medicine) for sharing their Dragonfly segmentation protocol; the imaging platforms ScopeM at ETH Zürich and the Center for Microscopy and Image Analysis at the University of Zurich for instrument access; and J. Döhner and J. Hehl for advice on STED imaging. T.R.-O. thanks L. Pignaton for assistance with Extended Data Fig. 1. C.S. was supported by the European Research Council (AdG TACKLE, 695192) and the Austrian Science Fund (FWF): W1257. M.P. was supported by the European Research Council (CoG CryoET4Diversity, 101000232) and the NOMIS Foundation.

Author information

Author notes

1. These authors contributed equally: Thiago Rodrigues-Oliveira, Florian Wollweber, Rafael I. Ponce-Toledo

Authors and Affiliations

1. Department of Functional and Evolutionary Ecology, Archaea Biology and Ecogenomics Unit, University of Vienna, Vienna, Austria

   Thiago Rodrigues-Oliveira, Rafael I. Ponce-Toledo, Simon K.-M. R. Rittmann & Christa Schleper

2. Institute of Molecular Biology & Biophysics, ETH Zürich, Zürich, Switzerland

   Florian Wollweber, Jingwei Xu & Martin Pilhofer

3. Plant Development & Electron Microscopy, Biocenter, Ludwig-Maximilans-Universität München, Planegg-Martinsried, Germany

   Andreas Klingl

Contributions

T.R.-O. performed all cultivation-related experiments, as well as qPCR and RT–qPCR and FISH analyses. R.I.P.-T. assembled and analysed the genome and performed phylogenetic calculations and comparative genomic analyses. F.W. acquired, processed and analysed cryo-ET data, performed western blotting and immunofluorescence experiments. F.W. and J.X. developed the ribosome-identification approach. J.X. performed sub-tomogram averaging of the ribosome and developed the filament analysis workflow. A.K. performed SEM and TEM analyses. S.K.-M.R.R. advised on anaerobic cultivation. C.S. collected the original sediment cores and water samples. C.S. and M.P. were responsible for the conceptualization and supervision of

the project. T.R.-O., F.W., R.I.P.-T., M.P. and C.S. wrote the manuscript with input from all of the authors.

Corresponding authors

Correspondence to Martin Pilhofer or Christa Schleper.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Jan Löwe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Microbial diversity profile of environmental samples (different brackish sediment core depths) and selected *L. ossiferum* enrichments.

Diversity estimations were made by amplicon sequencing using general prokaryotic 16S rRNA gene targeting primers (515f-806r). Taxonomic assignment was performed according to the SILVA database. Throughout the

enrichment process, the community became more homogeneous, leading to higher lokiarchaeal relative abundances.

Extended Data Fig. 2 Lokiarchaea distribution at different depths in a brackish sediment core sampled in Piran, Slovenia.

Abundance was estimated by submitting DNA extracted from different depths to qPCR assays using primers targeting lokiarchaeal 16S rRNA genes. Criteria for selecting good candidate fractions for enrichments were the number of copies in relation to DNA content. The sample with the highest lokiarchaeal 16S rRNA gene copies/g ratio to DNA content was the 13–16 cm fraction, which was then selected as inoculum for cultivation.

Extended Data Fig. 3 Specificity of qPCR primers LkF and LkR to Asgard archaea 16S rRNA genes.

Amplicons were generated from environmental DNA and submitted to Illumina Miseq (300 PE) sequencing. SILVA was used for taxonomic classification. These analyses show that the LkF and LkR primers designed in this study are highly specific to Asgard archaea 16S rRNA genes.

Extended Data Fig. 4 Functional annotation of groups of orthologous proteins encoded in *L. ossiferum* and '*Ca.* P. syntrophicum' genomes.

The venn diagram shows to scale the number of shared clusters of orthologous proteins between *L. ossiferum* and '*Ca.* P. syntrophicum' as well as genome-specific clusters. Numbers in parentheses indicate the number of clusters with no functional annotation. The graph shows that *L. ossiferum* is enriched in almost all categories compared to '*Ca.* P. syntrophicum', as expected for the larger genome size (>1.6 Mb difference). Functional categories: J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; D, cell cycle control, cell division and chromosome partitioning; V, defence mechanisms; T, signal transduction mechanisms; M, biogenesis of the cell wall, membrane or envelope; N, cell motility; U, intracellular trafficking, secretion and

vesicular transport; O, posttranslational modification, protein turnover and chaperones; X, mobilome; C, energy production and conversion; G, carbohydrate transport and metabolism; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism. For visualization purposes, clusters of orthologous proteins assigned to R (general function prediction only) and S (function unknown) categories are not shown. The numbers of clusters with only R and/or S assignment are: only in *L. ossiferum*, 405; shared, 617; only in '*Ca.* P. syntrophicum', 384.

Extended Data Fig. 5 Ideogram representation of syntenic regions (grey) between the genomes of *L. ossiferum* and '*Ca.* P. syntrophicum'.

Note that the genomes are represented in equal sizes (yellow and blue bars), although *L. ossiferum* has a 1.6 Mb larger genome.

Extended Data Fig. 6 Sub-tomogram averaging of the ribosome and identification of rRNA expansion segments.

a. Flowchart for the sub-tomogram averaging strategy of the ribosome. For details see Methods. **b**. Gold-standard FSC curve of the sub-tomogram average. **c**. Scheme of the ribosome identification workflow. For details see Fig. 3 and Methods. **d**. Multiple sequence alignment of expansion segments (ESs) 9 and 39 of the LSU rRNA of *L. ossiferum* compared to sequences from *T. kodakarensis* and those present in the MAGs of the organisms in the enrichment culture (Desulfovibrio_bin1: 44 contigs, 100% complete, 0% contamination; Methanogenium_bin1: 56 contigs, 98.7% complete, 2.6 % contamination; Methanogenium_bin2: 37 contigs, 98.7% complete, 1.9% contamination). The alignment reveals that ESs are only present in *L. ossiferum*.

Extended Data Fig. 7 Potential cell-cell interactions, variability of cell shape and rare vesicles observed in cryo-tomograms.

a. FISH analysis of the enrichment cultures with DAPI (cyan) and nucleotide probes targeting Lokiarchaea (red) and bacteria (green) and the corresponding phase contrast image showing isolated *L. ossiferum* cells. The FISH experiments were performed five independent times, with similar results. Bar: 2 µm. **b**/**c**. Aggregates of cells were infrequently observed during imaging of the enrichment culture. FISH analysis (b) with DAPI (cyan) and nucleotide probes targeting Lokiarchaea (red), bacteria (green) and *Methanomicrobiales* (purple) shows potential interactions of cells from different species. A low-magnification 2D EM image (c) shows *L. ossiferum* cells (Lo) next to a bacterial cell (B). Since the core of the cell assembly exceeded the thickness limitation of cryoET, no cryo-tomograms could be collected. Bars: 2 µm (b), 1 µm (c). **d**/**e**. Representative SEM (d) or low-magnification cryo-TEM images (e) showing different morphologies of *L. ossiferum* cells (n = 2 independent cultures). Both datasets revealed individual cell bodies that were always connected to at least one longer protrusion or vesicular structure. White arrowheads indicate the position of intracytoplasmic vesicles shown in f. Bars: 500 nm (d), 1 µm (e). **f**. Slices of cryo-tomograms (slice thickness 21.42 nm) showing the two intracytoplasmic vesicles observed in *L. ossiferum* protrusions (left) or a cell body (right), indicated by white arrowheads. Both vesicles seem fully enclosed by a membrane and show a low-density lumen. Bars: 100 nm.

Extended Data Fig. 8 Presence/absence pattern of chemotaxis-related proteins in Asgardarchaeota genomes.

The chemotaxis-related proteins encoded in *L. ossiferum* (Supplementary Table 4) together with the chemotaxis protein CheV MBN1502358.1 and the chemotaxis accessory protein CheZ CK5345549.1 were used as queries in BLASTp searches (evalue: $1e^{-10}$) against our Asgardarchaeota protein database followed by the annotation of putative chemotaxis proteins using BlastKOALA[87]. The presence/absence of proteins were mapped to the expanded phylogenetic tree shown in Fig. 2e. *L. ossiferum* is highlighted in red.

Extended Data Fig. 9 Analysis of cytoskeletal filaments by helical reconstruction of 2D-projected particles and sub-

tomogram averaging.

a. Flowchart of the *in situ* reconstruction of the cytoskeletal filament. For details see Methods. **b**. 2D classification of 2D-projected particles after different iterations of the reconstruction workflow and 2D classes obtained after the last iteration without sampling. **c**. Gold-standard FSC curve of the reconstruction after sub-tomogram averaging.

Extended Data Fig. 10 Immunogold localization of Lokiactin in *L. ossiferum*.

a. Conventional TEM micrograph of a *L. ossiferum* cell with no anti-Lokiactin antibodies used. Experiments were performed with samples derived from n = 2 independent cultures, with similar results. Bar: 500 nm. **b**. Micrographs of different *L. ossiferum* cells immunogold labelled with a specific anti-Lokiactin primary antibody (ab1, see Methods). Orange arrowheads point out gold beads. Experiments were performed three independent times, with similar results (n = 3). Bars: 500 nm. See also Supplementary Table 10.

Supplementary information

Supplementary Fig. 1

Uncropped western blot shown in Fig. 5f.

Reporting Summary

Peer Review File

Supplementary Tables

Supplementary Tables 1–10.

Supplementary Video 1

Cyo-electron tomograms and segmentations of a *L. ossiferum* protrusion network and a cell body.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Rodrigues-Oliveira, T., Wollweber, F., Ponce-Toledo, R.I. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* **613**, 332–339 (2023). https://doi.org/10.1038/s41586-022-05550-y

- Received: 13 June 2022
- Accepted: 10 November 2022
- Published: 21 December 2022
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05550-y

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

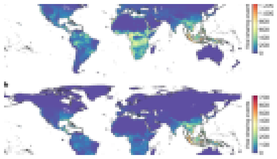
Provided by the Springer Nature SharedIt content-sharing initiative

Mysterious Asgard archaea microbes reveal their inner secrets

- Jan Löwe

Nature News & Views 21 Dec 2022

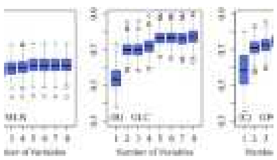
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05550-y



Climate change increases cross-species viral transmission risk

28 April 2022

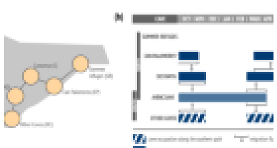
Colin J. Carlson, Gregory F. Albery, … Shweta Bansal



A Predictive Spatial Distribution Framework for Filovirus-Infected Bats

22 May 2018

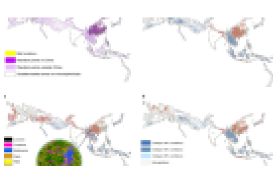
Graziano Fiorillo, Paolo Bocchini & Javier Buceta



Mechanisms for lyssavirus persistence in non-synanthropic bats in Europe: insights from a modeling study

24 January 2019

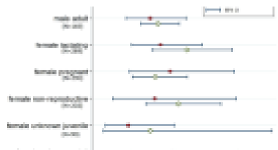
Davide Colombi, Jordi Serra-Cobo, … Vittoria Colizza



[Land-use change and the livestock revolution increase the risk of zoonotic coronavirus transmission from rhinolophid bats]

31 May 2021

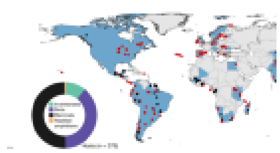
Maria Cristina Rulli, Paolo D’Odorico, … David T. S. Hayman



[Between roost contact is essential for maintenance of European bat lyssavirus type-2 in Myotis daubentonii bat reservoir: ‘The Swarming Hypothesis’]

03 February 2020

Daniel L. Horton, Andrew C. Breed, … Anthony R. Fooks



[Zoonotic host diversity increases in human-dominated ecosystems]

05 August 2020

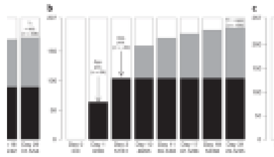
Rory Gibb, David W. Redding, … Kate E. Jones



[The land use–food–coronavirus nexus]

31 May 2021

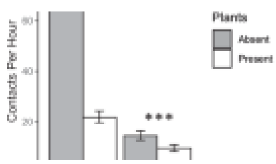
Monia Santini



[**Fluorescent biomarkers demonstrate prospects for spreadable vaccines to control disease transmission in wild bats**]

18 November 2019

Kevin M. Bakker, Tonie E. Rocke, … Daniel G. Streicker



[**Host density and habitat structure influence host contact rates and Batrachochytrium salamandrivorans transmission**]

27 March 2020

Daniel A. Malagon, Luis A. Melara, … Matthew J. Gray

- Article
- Open Access
- Published: 16 November 2022

Pathogen spillover driven by rapid changes in bat ecology

- Peggy Eby,
- Alison J. Peel,

- Andrew Hoegh,
- Wyatt Madden,
- John R. Giles,
- Peter J. Hudson &
- …
- Raina K. Plowright

Show authors

Nature **volume 613**, pages 340–344 (2023)

- 10k Accesses
- 1 Citations
- 1540 Altmetric
- Metrics details

Abstract

During recent decades, pathogens that originated in bats have become an increasing public health concern. A major challenge is to identify how those pathogens spill over into human populations to generate a pandemic threat[1]. Many correlational studies associate spillover with changes in land use or other anthropogenic stressors[2,3], although the mechanisms underlying the observed correlations have not been identified[4]. One limitation is the lack of spatially and temporally explicit data on multiple spillovers, and on the connections among spillovers, reservoir host ecology and behaviour and viral dynamics. We present 25 years of data on land-use change, bat behaviour and spillover of Hendra virus from Pteropodid bats to horses in subtropical Australia. These data show that bats are responding to environmental change by persistently adopting behaviours that were previously transient responses to nutritional stress. Interactions between land-use change and climate now lead to persistent bat residency in agricultural areas, where periodic food shortages drive clusters of spillovers. Pulses of winter flowering of trees in remnant forests appeared to prevent

spillover. We developed integrative Bayesian network models based on these phenomena that accurately predicted the presence or absence of clusters of spillovers in each of the 25 years. Our long-term study identifies the mechanistic connections between habitat loss, climate and increased spillover risk. It provides a framework for examining causes of bat virus spillover and for developing ecological countermeasures to prevent pandemics.

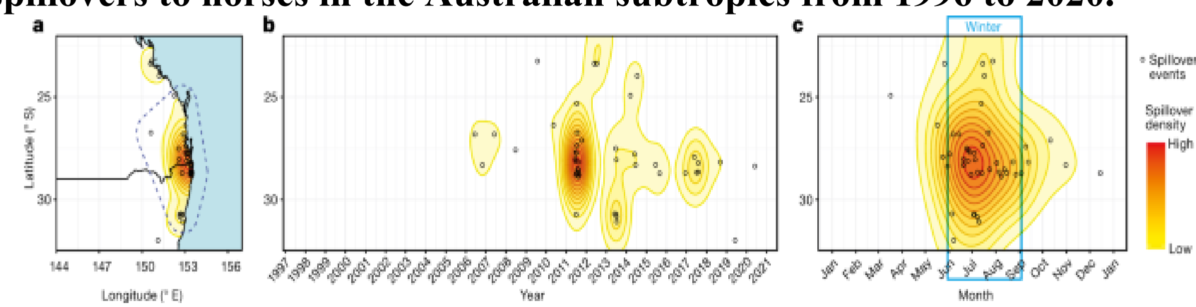
Main

Zoonotic spillover is the transmission of a pathogen from a non-human vertebrate to a human[1]. Spillovers of viruses from bats have resulted in the emergence and spread of viruses in the human population. For example, SARS-CoV-2, SARS-CoV-1, Nipah and Hendra viruses have caused human mortalities, sometimes after transmission through an intermediate host[5]. Spillover of viruses from wildlife to humans has been correlated with land-use change through studies that associate land use, occurrence of spillover and presence of reservoir hosts, but without data that reveal the mechanisms[1]. In this long-term study, we observed rapid changes in bat behaviour that coincided with the emergence of Hendra virus. We found that bats were responding to environmental change by persistently behaving in ways that were previously observed as temporary responses to climate-driven food shortages. We propose that these behavioural shifts increased spillover risk by increasing contact of bats with domestic horses, the intermediate hosts from which Hendra virus spills over into humans, and by increasing viral shedding from bat populations that have established outside their normal winter range[6]. We developed and applied a Bayesian hierarchical network model to our 25 years of data on reservoir host ecology, behaviour and spillover events (Data Index in ref. [7]). We identified distal and proximal drivers of links between habitat loss, climate and spillover, and predicted the risk of Hendra virus spillover in space and time. The Hendra virus system illustrates a suite of ecological connections that contribute to land-use-induced spillover of this pathogen: interactions between land-use change and climate altered the behaviour of wildlife reservoir hosts, increasing their proximity to domestic or human recipient hosts. We propose

that this phenomenon, coupled with stressors that drive increased pathogen excretion[6], leads to pathogen spillover.

Australian flying foxes (*Pteropus* spp., fruit bats) are reservoir hosts of Hendra virus, a henipavirus in the family Paramyxoviridae that does not cause discernible disease in bats, but has a case fatality rate of 75% in horses (84 fatalities documented) and 57% in humans (four fatalities documented)[8]. Among subtropical Pteropodids, *Pteropus alecto* (black flying fox) is the species most likely to excrete Hendra virus[9]. Infected bats feeding in horse paddocks shed the virus in excreta or spats that horses contact when grazing, leading to infection[9,10]. Humans are exposed through infected horses[8]. Hendra virus has probably circulated in bats far longer than Europeans have occupied Australia[11], yet Hendra spillover was not identified until 1994. After approximately 2006, the frequency of Hendra virus spillovers increased[10]. Forty one of the 63 spillovers documented to January 2021 occurred in subtropical eastern Australia[12]. The majority of spillovers occurred during the Austral winter and were clustered in space and time (Fig. 1 and Supplementary Information Section 1).

Fig. 1: Temporal and spatial distribution of documented Hendra virus spillovers to horses in the Australian subtropics from 1996 to 2020.



a, Distribution of spillovers across subtropical eastern Australia from 1996 to present. The dotted blue line denotes our study area. **b**, Hendra virus spillover by date of horse fatality from 1996 to present. Spillover events cluster in space in some years. **c**, Hendra virus spillover by month. In the subtropics, spillovers occur primarily during the Austral winter. The base-map was obtained from the Australian Bureau of Statistics digital boundary files (www.abs.gov.au).

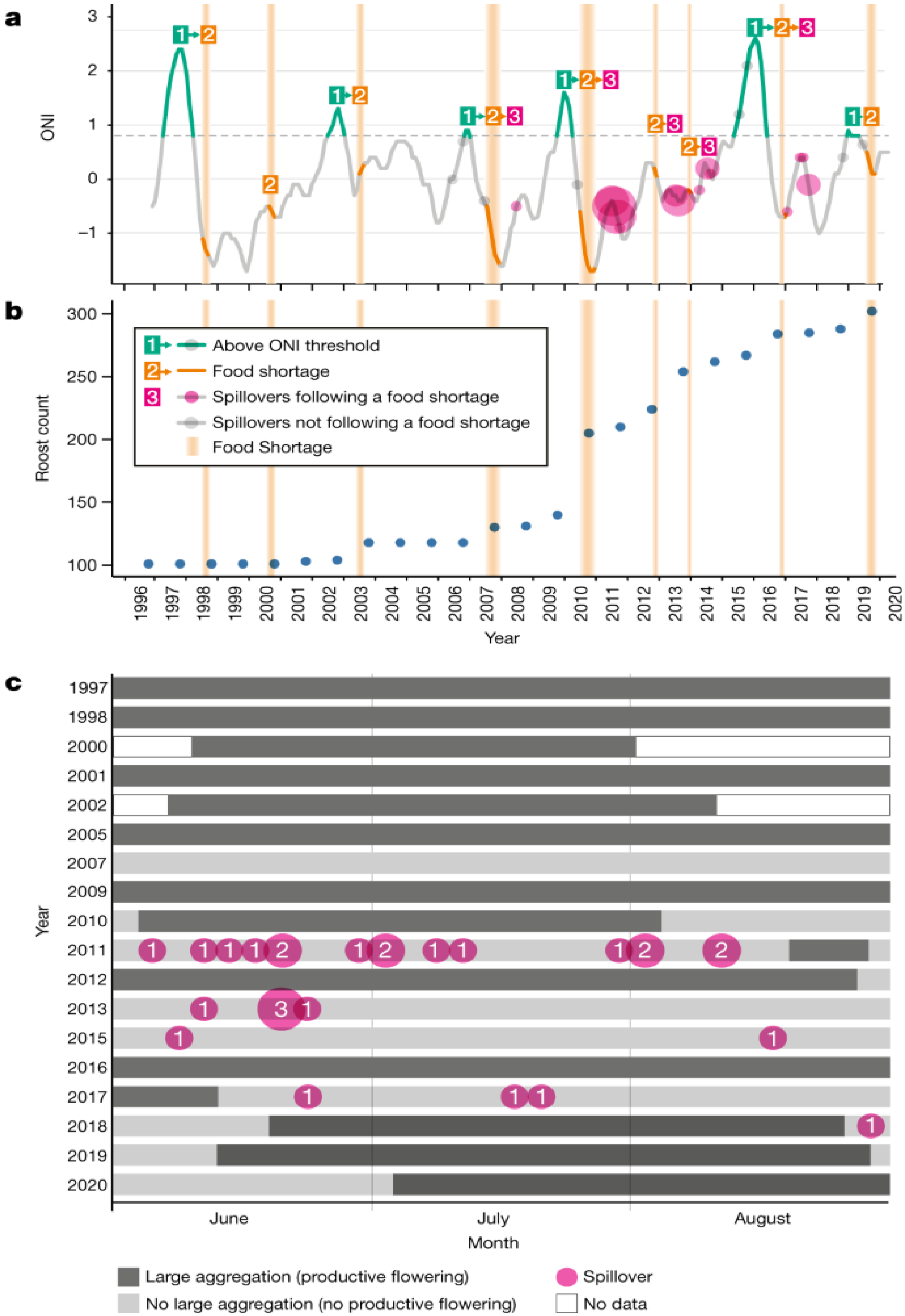
Our research focused on spillover to horses in the Australian subtropics, where *Pteropus* spp. primarily feed on nectar. Historically, the bats were nomadic, moving among roost sites to track ephemeral pulses of flowering by *Eucalyptus* spp. over hundreds of kilometres[13,14,15]; continuous occupancy of roosts was uncommon[13]. During summer, many tree species that are abundant and widely distributed provide food for bats[16]. During winter, few tree species provide food for bats, and the naturally limited distribution and abundance of these trees[16] has been further restricted by clearing for urban development and agriculture[17,18]. Pulses of mass flowering in remnant forests leads to a high proportion of the bat population becoming concentrated in small areas, typically in coastal lowlands[13]. Loss of these winter food sources can have severe impacts on the bat population[16,17]. When trees that provide food in winter or spring do not flower, as occurs every 1 to 4 years owing to variation in temperature and rainfall[19,20], bats experience brief (typically 3 to 12 week) food shortages. They respond to the shortages by roosting in small groups (population fission) close to reliable but often suboptimal food in urban gardens and agricultural areas (for example, fruit from ornamental, commercial or weed species)[13,17,21,22,23]. In the past, these behavioural responses persisted only for the duration of acute food stress, and bats returned to nomadism and nectivory when the food shortage abated[22,24,25].

To investigate the relationships between land-use change, changes in bat behaviour and spillover, we collected empirical data from 1996 to 2020 within an area bounded by the locations of spillovers in the subtropics (Fig. 1a, Extended Data Fig. 1a and Supplementary Information Section 1). These data include the location and timing of Hendra virus spillovers (Dataset A in ref. 12), locations and occupancy of the roosts (Datasets B,C)[26,27], characteristics of the roosts and foraging areas around the roosts (Datasets B,C,I)[26,27,28], climate (Dataset D in the Data Index)[7], nectar shortages (Dataset E)[29], measures of bat fitness (Dataset F and G)[30,31], flowering pulses during winter (Dataset J)[32] and loss of winter foraging habitat (Datasets K,L and M in the Data Index[7] and Supplementary Sections 1–13).

From 1996 until approximately 2002, roosting and foraging behaviours were stable, and no Hendra virus spillovers were detected (Figs. 1 and 2a,b). From

approximately 2003 until 2020, bat behaviour and the incidence of spillovers changed rapidly: the number of roosts tripled, and 40 spillovers were detected (Supplementary Information Sections 1 and 3 and Figs. 1 and 2). From 2006 to 2020, spillovers were detected in 80% of years, and 75% of spillovers occurred in annual clusters of three or more (Fig. 1). Neither a change in the definition of a Hendra virus case in horses in 2008, nor the availability of a horse vaccine in late 2012, explain the observed pattern of spillover events (Supplementary Information Section 1).

Fig. 2: Relationships between climate, periods of nutritional and energetic stress for bats and Hendra virus spillovers.



a
ONI
b
Roost count
Above ONI threshold
Food shortage
Spillovers following a food shortage
Spillovers not following a food shortage
Food Shortage
Year
c
Year
June
July
August
Month
Large aggregation (productive flowering)
No large aggregation (no productive flowering)
Spillover
No data

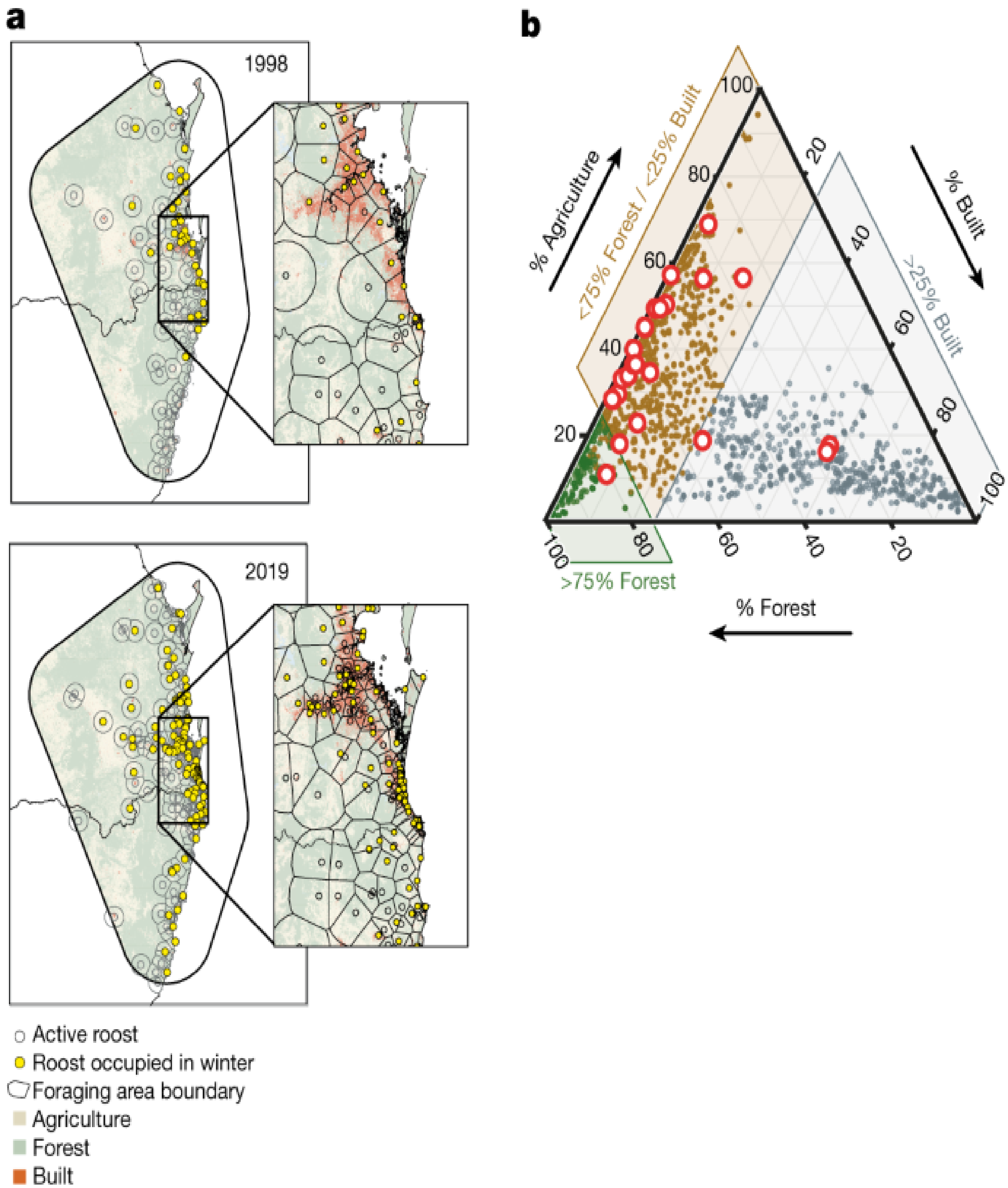
a, Temporal associations among the ONI, acute food shortages (identified by apiarists) and winter spillovers. A strong El Niño event (ONI > 0.8) consistently led to food shortages the following winter or spring, although food shortages can occur independently of ONI. During the early study period (1996 to 2002), food shortages did not lead to Hendra virus spillovers; during the period of rapid change, when bat populations were fissioning into urban and agricultural areas, food shortages led to Hendra virus spillovers during the following winter. **b**, Number of roosts in the study area. **c**, Timing of winter spillover events in relation to the presence of large aggregations (at least 100,000 bats) and productive flowering of diet species in southeast Australia during winter (June to August). Each row indicates a winter for which data were available. Data are missing for winter aggregations for 1999, 2003, 2004, 2006, 2008 and 2014 (Supplementary Table 5a).

We used nectar productivity data from commercial apiarists and measures of flying fox fitness to identify nine winter and spring food shortages during the study period (Supplementary Information Section 7 and Fig. 2a,b). During food shortages, the absence of nectar was associated with a sharp increase in the number of bats admitted into wildlife rehabilitation centres (at least 30 animals in a month) and a low percentage of lactating females with pre-weaning young (<79%; Supplementary Information Section 7 (ref. [33]) and Extended Data Figs. 2 and 3). Food shortages followed all strong El Niño events (Oceanic Niño Index ≥ 0.8), and also occurred independently of El Niño events (Fig. 2a).

Early in the study period, the number of roosts was stable. Although food shortages occurred, the bat population fission events were brief (weeks), bats quickly returned to nomadism when the shortage abated and the fission events were not captured in the roost counts. Following this, the number of roosts increased in steps that coincided with food shortages (2003, 2007 and 2010) and then continued to increase steadily (Fig. 2b). As the number of roosts increased, bats roosted in smaller groups that were closer together and fed within smaller areas, and new roosts formed in locations that increased access to anthropogenic foods in agricultural and urban areas (Fig. 3 and Extended Data Figs. 4 and 5a–c). We infer that fissioning of bat populations into smaller groups near anthropogenic food reduces the energetic costs of foraging and allows the bats to mitigate effects of nutritional and energetic

stress[25]. Whereas in the past, the behavioural response of bats to food shortages (population fission) persisted only for the duration of acute food stress[22,24,25], we observed the behavioural response becoming persistent (Extended Data Fig. 4). To assess the mechanisms associated with the rapid change in bat behaviour, we examined loss of winter foraging habitat across far Southeast Queensland (Supplementary Information 12 and Extended Data Fig. 1b). Before European settlement, winter habitat was extensive in this region (Extended Data Fig. 5d). More than 70% of the forest that provided winter habitat was cleared before 1996, and clearing continued at a constant rate (Extended Data Fig. 5e,g,h). In four of the six years from 1996 to approximately 2002, during which bat behaviour was consistent and no spillover occurred, we recorded aggregations of at least 100,000 nomadic bats associated with mass pulses of flowering in winter habitat in this region (Extended Data Fig. 5f). By 2018, nearly a third of the habitat that was present in 1996 had been cleared (Extended Data Fig. 5e); the number of roost sites had increased fivefold (Extended Data Fig. 4), with 87% of new roosts forming in urban areas. By 2020, winter aggregations of at least 100,000 in far Southeast Queensland had been recorded once in 14 years (Extended Data Fig. 5f). Although bats relocated to both urban and agricultural areas, most spillovers (86%) occurred in agricultural areas (Fig. 3b), presumably because horses were present at higher densities in agricultural areas than in urban areas.

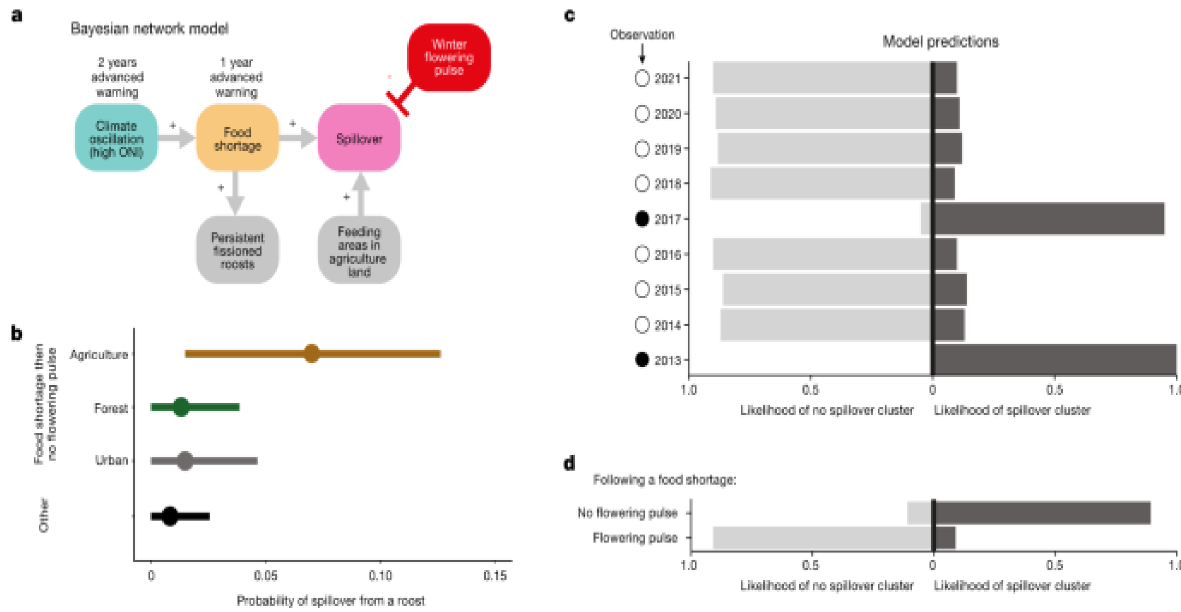
Fig. 3: The distribution of *Pteropus alecto* roosts during winter.



a, Expansion of the number and distribution of roosts occupied by *Pteropus alecto* during winter from 1998 to 2019. The base-map was obtained from the Australian Bureau of Statistics digital boundary files (www.abs.gov.au). **b**, Proportion of the foraging area surrounding roosts that was classified as built, forested or agricultural (Supplementary Information 13). The red circles indicate roosts that were the sources of winter spillovers.

We developed the parsimonious hypothesis that the combination of acute winter/spring food shortages followed by a lack of winter flowering in the following year yields the highest risk of spillover, especially from roosts in agricultural areas. We tested this mechanistic hypothesis with a multiscale Bayesian network model[34]. We included six variables: strong El Niño events (Oceanic Niño Index ≥ 0.8), food shortages, winter-flowering pulses, land cover within the foraging areas of winter colonies of *P. alecto*, population fissioning (establishment of new roosts) and number of Hendra virus spillovers (Supplementary Information Section 13 and Supplementary Fig. 1). We validated the predictive ability of the network model with leave-one-out cross-validation and expected log pointwise predictive density. We verified this by using a leave-one-out cross-validation for model selection with more than 1,000 simulated datasets (Supplementary Information Section 13). The best model supported that spillover clusters occur following a sequence of events over three successive years. First, it indicated that strong El Niño events preceded food shortages, and food shortages coincided with population fissioning (Fig. 4a, Extended Data Figs. 6–8 and Supplementary Tables 7–9). Second, the greatest likelihood of spillover from a given roost was associated with *P. alecto* feeding in agricultural areas during a winter that followed an acute food shortage, but in which there was no pulse of flowering (0.014, 0.126 highest posterior density probability interval; Fig. 4b). By contrast, the presence of a pulse of winter flowering that attracted at least 100,000 bats mitigated spillover risk (0.000, 0.026 probability interval; Fig. 2c and Fig. 4b, Extended Data Fig. 9 and Supplementary Table 5).

Fig. 4: Prediction of spillover on the basis of bat ecology and ecological conditions.



a, Structure of the Bayesian network model: strong El Niño events preceded food shortages that were associated with population fissioning and formation of persistent roosts in agricultural and urban areas. Spillover risk was greatest when *Pteropus alecto* fed in agricultural areas during a winter that followed an acute food shortage. The presence of a pulse of winter flowering that attracted large aggregations of bats mitigated spillover risk. Time delays between high ONI, food shortages and spillover allow advanced prediction, although winter-flowering pulses that mitigate spillover could not be predicted in advance. **b**, Probability of a spillover at the roost level given scenarios of land use under the condition of a food shortage followed by no winter-flowering pulse. Other, collates all other scenarios of the presence and absence of food shortages and flowering pulses. The circles are maximum a posteriori point estimates, and the bars are 95% highest posterior density credible intervals. **c**, Predicted probability of a cluster of spillovers in each year from 2013 to 2021. Predictions for a given year were made with all observations to that date; information from future years was not incorporated into predictions. In 2017 and 2013, food shortages in the year before and no winter-flowering pulse led to elevated risk of spillover, whereas in the other years, no clusters of spillovers were predicted. All predictions were consistent with the realized data on the absence (open circles) or presence (closed circles) of clusters of spillover. **d**, Predictive probability of a cluster of spillovers in winter 2020, following a food shortage in 2019. A winter-flowering pulse occurred in early July 2020 and

no winter spillovers were observed. We predicted that a cluster of spillover events would have occurred if there had not been a flowering pulse.

We used the model to make probabilistic predictions of whether a cluster of spillovers (defined as three or more spillovers) would occur in the study area from 1996 to 2021. We used all available information up to a given time point, starting with 2013 (Fig. 4c and Extended Data Fig. 10a,b). For example, to predict a cluster of spillovers in 2020 (Fig. 4d and Extended Data Fig. 10c), we used information collected to the end of 2019. In all cases, model predictions were consistent with observations of the presence or absence of clusters of spillovers in the context of winter flowering. We also verified the predictive ability of our model with fivefold cross-validation; our model accurately predicted the presence or absence of a cluster of spillovers in each of the 25 years in the data (Extended Data Fig. 8).

These data document increasing rates of transmission of a fatal zoonotic virus in a rapidly changing anthropogenic system. Our work suggests that previously transient fissioning of bat populations in response to periods of food shortage has become persistent. This behavioural response leads to the increased use of agricultural and urban areas by bats in winter (Fig. 3). Food shortages occurred throughout the 25-year study period (Fig. 2a,b), probably driven by the lagged effects of climate on *Eucalyptus* flowering[20]. However, spillovers were not associated with food shortages until bats persistently overwintered in agricultural areas in proximity to horses, and winter-flowering pulses became less frequent (Figs. 2c,3 and 4a). We propose that the loss of winter-flowering habitat and consequent decline in the abundance of winter nectar contributes to the persistence of bats in agricultural and urban areas (Supplementary Information Section 12, Fig. 2c and Extended Data Fig. 5d–h). The consequences of more bats in areas with human settlements include not only increased risk of viral spillover from bats to horses to humans, but also increased conflict with humans[35].

The contemporary association between food shortages, population fission and spillover may be explained by both increased proximity of bats and horses, and physiological stress in novel winter habitats. Bats in roosts that recently formed outside of the historic winter range of the species excreted more Hendra virus in winter, especially after food shortages, than bats in

roosts within the historic winter range of the species[6]. Bats in urban and agricultural areas rely on suboptimal food, which may lead to nutritional stress that facilitates viral shedding[36,37,38,39]. Nutritional stress has been linked to infection and shedding of Hendra virus[36,39], coronaviruses[40] and other viruses in reservoir host systems[41]; moreover, Hendra virus spillover events in the subtropics have been associated with small roosts with limited access to native food[19]. The timing of Hendra virus spillover clusters in winter, months after the food shortages in the previous year, may be owing to the cumulative effects of nutritional stress overlaying high energy requirements in winter (thermoregulation and pregnancy) and scarce resources within suboptimal habitats. Although the time-lag between food shortages and spillover requires further investigation, we suggest that the processes driving spillover of Hendra virus are general phenomena that link land-use change and pathogen emergence: behavioural responses to loss of habitat increase contact with recipient hosts, and physiological and immunological responses to food limitation increase pathogen excretion[4].

Our data suggest that increasingly rare winter-flowering pulses reduce the risk of spillover. Bats reverted to nomadism and left agricultural and urban areas during pulses of winter flowering in remnant native forest, and spillovers did not occur during these flowering pulses (Fig. 2c). We propose that these pulses of flowering may mitigate zoonotic risk by drawing large numbers of bats (Supplementary Information Section 11 (ref. [32])) away from feeding in agricultural areas and, therefore, decreasing contact between bats and horses. This nomadism may also reduce competition for food among the bats that remain in agricultural areas. Understanding these mechanisms requires further work. Nevertheless, the loss of native forest that supports large aggregations of nomadic bats appears to be fundamental to the cascade of events that lead to spillover. An extensive programme of ecological protection and restoration of winter-flowering forests (ecological countermeasures) could be a sustainable, long-term strategy to reduce spillover and protect the health of livestock and humans[42].

The consistent temporal association of Oceanic Niño Index (ONI) thresholds with food shortages and spillover events allows for the prediction of spillover clusters up to two years in advance through surveillance of climate, or alternatively, one year in advance through surveillance of bat

reproduction and bat admissions to wildlife rehabilitation centres (Fig. 4c). However, we cannot as yet predict winter-flowering pulses, and these must be monitored in real time. For example, the conditions leading into winter 2020 suggested that multiple spillover events would occur (Fig. 4d). However, we were unable to predict the pulse of winter flowering that attracted more than 200,000 bats that appears to have prevented winter spillovers. With retrospective information on flowering, our Bayesian network model correctly predicted the absence of a cluster of spillover events in 2020.

Prevention of spillovers requires characterization of the interactions of reservoir and recipient species with their environment in the context of rapid land-use change and climate variability. Monitoring of ecological processes that occur at different temporal extents, including those longer than natural climatic cycles such as El Niño, is essential to understanding the drivers of spillover. Yet long-term data on reservoir hosts, especially bats, are sparse[43]. Many bat species that are the reservoir hosts of zoonotic pathogens depend on ephemeral resources, and occupy ecosystems in which loss of native vegetation, high livestock density and human populations coincide[44,45,46]. Moreover, some bat species can adapt to human-modified landscapes to mitigate the effects of loss of their native habitat. These species may present greater risks of spillover, as has been observed in other wildlife taxa[47]. We identified key processes connecting land-use change to spillover through behavioural responses of bats to altered food availability. We suggest that behavioural and physiological responses to rapid, human-induced environmental change increased contact between reservoir hosts and recipient hosts, and increased shedding of pathogen[6] within proximity of recipient hosts. This study, therefore, suggests a general framework for examining causes and potential ways to mitigate bat virus spillover in regions without long-term data on reservoir hosts.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets generated and analysed during the current study are available in the Cornell University eCommons Digital Repository or they are available as open access files. The URLs are provided in the Data Index https://doi.org/10.7298/pjjb-3360, with the exception of records from commercial apiarists (Supplementary Information Section 7) that are constrained by commercial in-confidence considerations. Dataset URLs: Dataset A, https://doi.org/10.7298/3dbp-t721; Dataset B, https://doi.org/10.7298/kdht-sp38; Dataset C, https://doi.org/10.7298/ajmw-mp18; Dataset E, https://doi.org/10.7298/tb5p-dr98; Dataset F, https://doi.org/10.7298/j3q2-gw32; Dataset G, https://doi.org/10.7298/3vha-5m37; Dataset I, https://doi.org/10.7298/x71e-c660; Dataset J, https://doi.org/10.7298/rmhz-dc23.

Code availability

Code generated during the current study, and the input data for the models, are available in the Supplementary Information and Cornell University eCommons Digital Repository: https://doi.org/10.7298/rbde-cy49; https://doi.org/10.7298/y0nr-e545.

References

1. Plowright, R. K. et al. Pathways to zoonotic spillover. *Nat. Rev. Microbiol.* **15**, 502–510 (2017).

2. Jones, K. E. et al. Global trends in emerging infectious diseases. *Nature* **451**, 990–993 (2008).

3. Gottdenker, N. L., Streicker, D. G., Faust, C. L. & Carroll, C. Anthropogenic land use change and infectious diseases: a review of the evidence. *Ecohealth* **11**, 619–632 (2014).

4. Plowright, R. K. et al. Land use-induced spillover: a call to action to safeguard environmental, animal, and human health. *Lancet Planet. Health* **5**, e237–e245 (2021).

5. Letko, M., Seifert, S. N., Olival, K. J., Plowright, R. K. & Munster, V. J. Bat-borne virus diversity, spillover and emergence. *Nat. Rev. Microbiol.* **18**, 461–471 (2020).

6. Becker, D. J., Eby, P., Madden, W., Peel, A. J. & Plowright, R. K. Ecological conditions predict the intensity of Hendra virus excretion over space and time from bat reservoir hosts. *Ecol. Lett.* https://doi.org/10.1111/ele.14007 (2022).

7. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Data Index. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/pjjb-3360 (2022).

8. Halpin, K. et al. Pteropid bats are confirmed as the reservoir hosts of henipaviruses: a comprehensive experimental study of virus transmission. *Am. J. Trop. Med. Hyg.* **85**, 946–951 (2011).

9. Edson, D. et al. Routes of Hendra virus excretion in naturally-infected flying-foxes: implications for viral transmission and spillover risk. *PLoS ONE* **10**, e0140670 (2015).

10. Plowright, R. K. et al. Ecological dynamics of emerging bat virus spillover. *Proc. Royal Soc. B.* **282**, 20142124 (2015).

11. Wang, L. F., Walker, P. J. & Poon, L. L. M. Mass extinctions, biodiversity and mitochondrial function: are bats 'special' as reservoirs for emerging viruses? *Curr. Opin. Virol.* **1**, 649–657 (2011).

12. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Dataset A: register of Hendra virus spillovers to horses. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/3dbp-t721 (2022).

13. Fleming, T. H. & Eby, P. in *Ecology of Bats* (eds Kunz, T. H. & Fenton, M. B.) 156–208 (Univ. of Chicago Press, 2003).

14. Eby, P. Seasonal movements of grey-headed flying-foxes, *Pteropus poliocephalus* (Chiroptera, Pteropodidae), from 2 maternity camps in Northern New-South-Wales. *Wildl. Res.* **18**, 547–559 (1991).

15. Roberts, B. J., Catterall, C. P., Eby, P. & Kanowski, J. Long-distance and frequent movements of the flying-fox *Pteropus poliocephalus*: implications for management. *PLoS ONE* **7**, e42532 (2012).

16. Eby, P. & Law, B. *Ranking the Feeding Habitat of Grey-headed Flying Foxes for Conservation Management* (Department of Environment and Climate Change (NSW) & Department of Environment, Water, Heritage, & the Arts, 2008).

17. Eby, P., Richards, G., Collins, L. & Parry-Jones, K. The distribution, abundance and vulnerability to population reduction of a nomadic nectarivore, the grey-headed flying-fox *Pteropus poliocephalus* in New South Wales, during a period of resource concentration. *Aust. Zool.* **31**, 240–253 (1999).

18. Catterall, C. P., Kingston, M. B., Park, K. & Sewell, S. Deforestation, urbanisation and seasonality: interacting effects on a regional bird assemblage. *Biol. Conserv.* **84**, 65–81 (1998).

19. Giles, J. R., Plowright, R. K., Eby, P., Peel, A. J. & McCallum, H. Models of eucalypt phenology predict bat population flux. *Ecol. Evol.* **6**, 7230–7245 (2016).

20. Hudson, I. L., Keatley, M. & Kang, I. Wavelet characterization of eucalypt flowering and the influence of climate. *Environ. Ecol. Stat.* **18**, 513–533 (2011).

21. Páez, D. J., Restif, O., Eby, P. & Plowright, R. K. Optimal foraging in seasonal environments: implications for residency of Australian flying foxes in food-subsidized urban landscapes. *Philos. Trans. R. Soc. B* **373**, 20170097 (2018).

22. Ratcliffe, F. N. *The Flying Fox in Australia (Pteropus).* Bulletin No. 53 (Commonwealth Council for Scientific and Industrial Research, 1931).

23. Meade, J., Martin, J. M. & Welbergen, J. A. Fast food in the city? Nomadic flying-foxes commute less and hang around for longer in urban areas. *Behav. Ecol.* **32**, 1151–1162 (2021).

24. Eby, P. *The Biology and Management of Flying-foxes in New South Wales*. Report No. 18 (New South Wales National Parks and Wildlife Service, 1995).

25. Wong, B. & Candolin, U. Behavioral responses to changing environments. *Behav. Ecol.* **26**, 665–673 (2015).

26. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Dataset B: register of flying fox roosts in the study area. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/kdht-sp38 (2022).

27. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Dataset C: SEQ monthly roost distribution and population estimates. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/ajmw-mp18 (2022).

28. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Dataset I: landcover data. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/x71e-c660 (2022).

29. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Dataset E: months of nectar shortage. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/tb5p-dr98 (2022).

30. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Dataset F: records of wildlife rehabilitation centers. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/j3q2-gw32 (2022).

31. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Dataset G: assessments of pre-weaning reproductive output. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/3vha-5m37 (2022).

32. Eby, P. et al. Data from pathogen spillover driven by rapid changes in bat ecology. Dataset J: winter flower pulses. *Cornell University*

eCommons Digital Repository https://doi.org/10.7298/rmhz-dc23 (2022).

33. Eby, P. et al. Data and scripts from pathogen spillover driven by rapid changes in bat ecology. Food shortage regression tree model. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/rdbe-cy49 (2022).

34. Eby, P. et al. Data and figure from pathogen spillover driven by rapid changes in bat ecology. Bayesian network model. *Cornell University eCommons Digital Repository* https://doi.org/10.7298/y0nr-e545 (2022).

35. Roberts, B. J., Mo, M., Roache, M. & Eby, P. Review of dispersal attempts at flying-fox camps in Australia. *Aust. J. Zool.* **68**, 254–272 (2020).

36. Plowright, R. K. et al. Reproduction and nutritional stress are risk factors for Hendra virus infection in little red flying foxes (*Pteropus scapulatus*). *Proc. Royal Soc. B.* **275**, 861–869 (2008).

37. Plowright, R. K. et al. Transmission or within-host dynamics driving pulses of zoonotic viruses in reservoir–host populations. *PLoS Negl. Trop. Dis.* **10**, e0004796 (2016).

38. Peel, A. J. et al. Synchronous shedding of multiple bat paramyxoviruses coincides with peak periods of Hendra virus spillover. *Emerg. Microbes Infect.* **8**, 1314–1323 (2019).

39. Edson, D. et al. Time of year, age class and body condition predict Hendra virus infection in Australian black flying foxes (*Pteropus alecto*). *Epidemiol. Infect.* **147**, e240 (2019).

40. Wacharapluesadee, S. et al. Longitudinal study of age-specific pattern of coronavirus infection in Lyle's flying fox (*Pteropus lylei)* in Thailand. *Virol. J.* **15**, 38 (2018).

41. Owen, J. et al. Reservoir hosts experiencing food stress alter transmission dynamics for a zoonotic pathogen. *Proc. R. Soc. Lond. B*

Biol. Sci. **288**, 20210881 (2021).

42. Reaser, J. K., Witt, A., Tabor, G. M., Hudson, P. J. & Plowright, R. K. Ecological countermeasures for preventing zoonotic disease outbreaks: when ecological restoration is a human health imperative. *Restor. Ecol.* **29**, e13357 (2021).

43. Becker, D. J., Crowley, D. E., Washburne, A. D. & Plowright, R. K. Temporal and spatial limitations in global surveillance for bat filoviruses and henipaviruses. *Biol. Lett.* **15**, 20190423 (2019).

44. Kessler, M. K. et al. Changing resource landscapes and spillover of henipaviruses. *Ann. N.Y. Acad. Sci.* **1429**, 78 (2018).

45. McKee, C. D. et al. The ecology of Nipah virus in Bangladesh: a nexus of land-use change and opportunistic feeding behavior in bats. *Viruses* **13**, 169 (2021).

46. Rulli, M. C., D'Odorico, P., Galli, N. & Hayman, D. T. Land-use change and the livestock revolution increase the risk of zoonotic coronavirus transmission from rhinolophid bats. *Nat. Food* **2**, 409–416 (2021).

47. Gibb, R. et al. Zoonotic host diversity increases in human-dominated ecosystems. *Nature* **584**, 398–402 (2020).

Acknowledgements

We thank B. Han, H. McCallum, M. Ruiz-Aravena, T. Lunn, L. McGuire, E. Fleishman, G. Tabor, D. Kerlin, B. Hunt, M. Higgs and T. Creech for comments on the manuscript or statistical methods. We thank R. Egloff for help with the figures. We thank P. Freeman, P. Kemsley, L. Ruytenberg, J. Wood, L. Collins, S. Stanford, P. Space, B. White, T. Whitehead, B. Winner, D. Sommerville, P. Reid, H. Field, N. Kung and C. Smith for helpful conversations and data. The Griffith University Human Research Ethics Committee (GUHREC) approved the research (GU Ref No. 2022/765) and informed consent was obtained from all human research participants. This research was developed with funding from the National Science Foundation

(DEB-1716698), US Defense Advanced Research Projects Agency (DARPA PREEMPT D18AC00031) and US National Institute of Food and Agriculture (1015891). A.J.P. was supported by an Australian Research Council DECRA fellowship (DE190100710). The content of the information does not necessarily reflect the position or the policy of the US government, and no official endorsement should be inferred. We acknowledge the Biripi, Bundjalung, Butchulla, Dainggatti, Danggan Balun, Githabul, Gumbainggir, Kabi Kabi, Turrbal, Wakka Wakka, Widjabul Wia-bal, Yaegl, Yugambeh and Yuggera Ugarapul people, who are the Traditional Custodians of the land upon which this work was conducted.

Author information

Author notes

1. Wyatt Madden

   Present address: Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, USA

2. John R. Giles

   Present address: Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA

Authors and Affiliations

1. School of Biological, Earth, and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Australia

   Peggy Eby

2. Centre for Planetary Health and Food Security, Griffith University, Nathan, Queensland, Australia

   Peggy Eby & Alison J. Peel

3. Center for Large Landscape Conservation, Bozeman, MT, USA

Peggy Eby

4. Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA

Andrew Hoegh

5. Department of Microbiology and Cell Biology, Montana State University, Bozeman, MT, USA

Wyatt Madden & Raina K. Plowright

6. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

John R. Giles

7. Center for Infectious Disease Dynamics, Pennsylvania State University, State College, PA, USA

Peter J. Hudson

8. Department of Public and Ecosystem Health, Cornell University, Ithaca, NY, USA

Raina K. Plowright

Contributions

P.E. collected long-term empirical data and collated existing data. R.K.P., P.E. and A.J.P. wrote the first draft of the manuscript and P.E., R.K.P., A.J.P., P.J.H. and A.H. edited the manuscript. P.E., A.J.P., A.H., W.M. and R.K.P. wrote the Supplementary Information. A.H. conceptualized and implemented the Bayesian network model. All authors contributed to analysis of the data and production of the figures.

Corresponding author

Correspondence to Raina K. Plowright.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks DeeAnn Reeder and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Boundaries of the Hendra virus study area and Southeast Queensland study area with locations of Hendra virus spillovers to horses indicated.

(a) The full study area with locations of subtropical Hendra virus spillovers (circles, Database A)[12]; **(b)** The far Southeast Queensland (SEQ) study area showing the Hendra virus spillovers documented within the boundary (circles, Database A)[12]; and the 2018 distributions of built land use (light red, Supplementary Information 9) and extant winter foraging habitat for flying foxes (green, Supplementary Information 12). Base-maps were obtained from the Australian Bureau of Statistics Digital boundary files (www.abs.gov.au).

Extended Data Fig. 2 Measures of flying fox fitness identified as important proxies of food shortage in regression tree model.

(a) index of reproductive output: annual estimates of the percentage of adult females occupying roosts in the Northern Rivers (NR) region (light grey) and throughout the Hendra virus study area (black) carrying live young pre-weaning; data are grouped according to whether a period of nectar shortage was recorded by apiarists during the winter or spring associated with that birth cohort (n = 9 years) versus cohorts when a nectar shortage was not recorded (n = 14 years). Data acquired in years of food shortage were widely dispersed relative to highly clustered data acquired in years without food shortage. Patterns of data distribution were consistent between the NR region and the full study area. Box plots show median values, upper and lower quartiles and all data points. **(b)** monthly counts of admissions to Northern Rivers WIRES wildlife rehabilitation center with months classified as nectar shortage by apiary data (indicated by orange bars). No intake records were available from June 2003 through December 2005. The bounds of the box correspond to the 25th and 75th percentiles, the middle dash is the median, and the whiskers extend to the largest value no further than 1.5 times the interquartile range from the 25th or 75th percentile.

Extended Data Fig. 3 Regression tree fit for identifying flying fox food shortages (as assessed by absence or very low nectar production).

Food shortages were predicted when monthly flying fox intakes into wildlife rehabilitation centers were ≥30 individuals and minimum proportion of female flying foxes with young pre-weaning was <79%.

Extended Data Fig. 4 Changes through time in the number and estimated population size of roosts occupied by black flying foxes *Pteropus alecto* in the SEQ study area.

Censuses of roosts were conducted monthly in two time periods: T1 = 1996–1999 and T2 = 2009–2012. **(a)** Monthly maps of roost locations and population size grouped in three quantiles. The base-map was obtained from

the Australian Bureau of Statistics Digital boundary files (www.abs.gov.au). **(b)** An additional population size category (>25,000) was added to illustrate change in large roosts. The line is a LOESS best fit and gray shading indicates 95% confidence intervals. **(c)**. Total population estimates of black flying foxes (all roosts combined) grouped by month and year and summarized in violin plots. The number and locations of roosts was stable in T1 (n = 12) and increased from 34 to 80 in T2, although total estimates of black flying foxes in the study area remained stable.

Extended Data Fig. 5 Spatial and temporal changes in the roosting and foraging behaviors of *Pteropus* bats during the study period (1996–2019 comparisons); and coincident decline in the extent of winter habitat in the SEQ study area.

(a) Mean distance between nearest neighboring roosts and the distance between roosts and developed areas decreased. **(b)** Roost occupancy in built land in metropolitan centers increased; smaller foraging areas contained a higher proportion of urban land; **(c)** Annual change in the area of different land-cover types within the total winter foraging range of *P. alecto*. Data are the summed total of the area assigned to each land use type in the foraging areas of roosts occupied in winter. The total area of land in the winter foraging range increased over time, primarily as a result of expansion into agricultural areas. Although the number of winter roosts in urban areas increased substantially, the total area of built land in combined foraging areas did not. **(d)** Distribution of key winter habitat in the southeast Queensland study area (SEQ): pre-clearing, at the start of the study in 1996 and at the end of the study in 2018. **(e)** Percentage of pre-clearing winter habitat remaining (note that the y-axis is compressed). **(f)** Abundance of bats in the largest winter roosts within SEQ. Aggregations ≥100,000 were common pre-2005, rare after 2005, and none occurred from 2011 to 2020. **(g)** Annual rate of clearing of key winter habitat in the SEQ study area 1996–2018, the line is a LOESS best fit. Gray shading indicates 95% confidence intervals and **(h)** Cumulative area cleared (ha).

Extended Data Fig. 6 A total of 64 different plausible network models were assessed using expected log pointwise predictive

density (ELPD) and leave-one-out (LOO) cross-validation for predictive validation.

(a) The figure contains the best network, Model 36 (top), plus all of the other networks within 10 leave-one-out information criteria (LOOIC) units of Model 36. The node labels are abbreviated; spill is spillover, roost is roost fissioning, land use is the land use type at the roost level, short is a food shortage in the previous year, pulse is a winter flowering pulse, and ONI is elevated ONI over 0.8 two years prior. **(b)** Distribution of LOOIC values relative to Model 36. For each model, the boxplots (n = 1000 simulations) show the distribution of the difference in LOOIC. M36 is the favored model as the median, and at least the first quartile, are positive for all other models. The bounds of the box correspond to the 25^{th} and 75^{th} percentiles, the middle dash is the median, and the whiskers extend to the largest value no further than 1.5 times the interquartile range from the 25^{th} or 75^{th} percentile.

Extended Data Fig. 7 Posterior distributions for model parameters in Bayesian Network Model.

(A) posterior distribution of new fissioned roosts, conditional on food shortage events. **(B)** posterior distribution for the probability of a winter flower pulse. **(C)** posterior distribution for the probability of a roost having the three different land use characteristics. **(D)** posterior distribution for the probability of a food shortage given elevated ONI values. **(E)** the posterior distribution for the probability of elevated ONI values.

Extended Data Fig. 8 Output from 5-fold cross validation study to assess the ability of our model to predict spillover clusters.

The red bars correspond to whether a cluster of spillovers was observed in that year and the grey bars and number correspond to predicted probabilities. For a given year, roost-level predictions are made using models fit on roosts-year combinations from other folds. This approach provides strong evidence of the predictive ability of the model framework as it predicts, with high certainty, the presence or absence of clusters of spillovers in all 25 years.

Extended Data Fig. 9 The locations of pulses of winter flowering in southeast Australia 1997–2020 as identified by aggregations ≥100,000 of nomadic grey-headed flying foxes and black flying foxes; and the winter flowering diet plants that occur in local vegetation communities.

Winter habitat in the Hendra virus study area primarily comprises a mix of four species, two single species occur in other regions. Latitudinal variations in the flowering phenology of *Corymbia maculata* are indicated. Dates are provided in Dataset J[32], a key to roost names is provided in Supplementary Table 5B. The base-map was obtained from the Australian Bureau of Statistics Digital boundary files (www.abs.gov.au).

Extended Data Fig. 10 Predicted probabilities for total number and clusters of winter spillovers.

(a) Predicted probability for total number of winter spillovers (grey bars) and observed number of winter spillovers (red outlines). **(b)** Predicted probability for a cluster of spillovers (grey bars, defined as three or more spillovers) and the observed presence or absence of a cluster of winter spillovers (red outlines). In both figures, predictions are made using all observed data up to that time, but information from future years is not incorporated into predictions. The period of 2016 through 2021 was chosen for probabilistic predictions because enough preceding data allowed predictions. Food shortages were seen preceding years 2017 and 2020, but only the year 2017 included both a food shortage and a lack of a winter flowering pulse. All the predictions are supported by the realized data, which fall in higher mass parts of the distributions in **(a)** and align with predictions with high predicted probabilities in **(b)**. **(c)** Predictive distribution of the total number of spillovers in winter 2020, following a food shortage in 2019. The grey bars indicate the spectrum of probabilistic predictions for spillover, given a winter flowering pulse. A winter flowering pulse did occur in early July 2020 and no winter spillovers were observed (pink outlined bar). Had no flowering pulse occurred, we predicted many more spillover events (yellow bars).

Supplementary information

Supplementary Information

This file contains details of the methods used to generate and analyse data for this study and extended supporting material that assists with the interpretation of long-term datasets. The file includes links to a Data Index that provides an overview of the datasets used in the study, descriptions of data fields, data sources, availability of data, the analyses the datasets contributed to, and the Supplementary Information sections where those analyses are described. Links are also provided to access the compiled input data for models and their output. The file includes methods for the multiscale Bayesian network model, methods used to generate data for the model and context for interpreting the results.

Reporting Summary

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Eby, P., Peel, A.J., Hoegh, A. *et al.* Pathogen spillover driven by rapid changes in bat ecology. *Nature* **613**, 340–344 (2023). https://doi.org/10.1038/s41586-022-05506-2

- Received: 20 January 2022
- Accepted: 01 November 2022
- Published: 16 November 2022
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05506-2

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

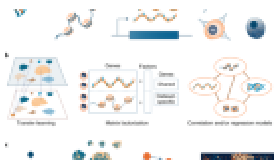
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05506-2



Automated brightfield morphometry of 3D organoid populations by OrganoSeg

28 March 2018

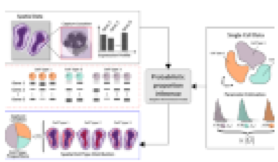
Michael A. Borten, Sameer S. Bajikar, … Kevin A. Janes



Computational methods for single-cell omics across modalities

06 January 2020

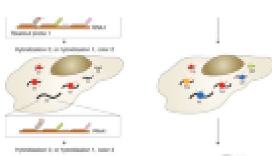
Mirjana Efremova & Sarah A. Teichmann



Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography

09 October 2020

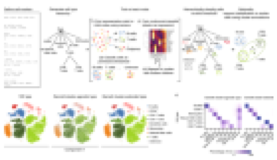
Alma Andersson, Joseph Bergenstråhle, … Joakim Lundeberg



Spatially resolved single-cell genomics and transcriptomics by imaging

06 January 2021

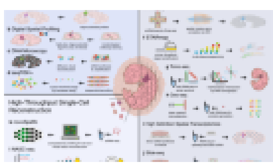
Xiaowei Zhuang



Supervised classification enables rapid annotation of cell atlases

09 September 2019

Hannah A. Pliner, Jay Shendure & Cole Trapnell



From whole-mount to single-cell spatial assessment of gene expression in 3D

23 October 2020

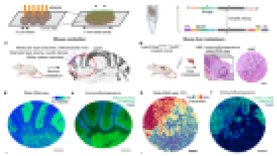
Lisa N. Waylen, Hieu T. Nim, … Mirana Ramialison



Phase contrast time-lapse microscopy datasets with automated and manual cell tracking annotations

13 November 2018

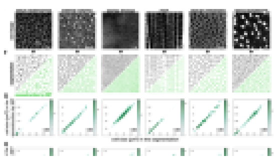
Dai Fei Elmer Ker, Sungeun Eom, … Phil G. Campbell



[Spatial genomics enables multi-modal study of clonal heterogeneity in tissues]

15 December 2021

Tongtong Zhao, Zachary D. Chiang, … Fei Chen



[Unraveling spatial cellular pattern by computational tissue shuffling]

23 October 2020

Elise Laruelle, Nathalie Spassky & Auguste Genovesio

- Article
- Open Access
- Published: 04 January 2023

Integrated intracellular organization and its variations in human iPS cells

- Matheus P. Viana,

- Jianxu Chen,
- Theo A. Knijnenburg,
- Ritvik Vasan,
- Calysta Yan,
- Joy E. Arakaki,
- Matte Bailey,
- Ben Berry,
- Antoine Borensztejn,
- Eva M. Brown,
- Sara Carlson,
- Julie A. Cass,
- Basudev Chaudhuri,
- Kimberly R. Cordes Metzler,
- Mackenzie E. Coston,
- Zach J. Crabtree,
- Steve Davidson,
- Colette M. DeLizo,
- Shailja Dhaka,
- Stephanie Q. Dinh,
- Thao P. Do,
- Justin Domingus,
- Rory M. Donovan-Maiye,
- Alexandra J. Ferrante,
- Tyler J. Foster,
- Christopher L. Frick,
- Griffin Fujioka,
- Margaret A. Fuqua,
- Jamie L. Gehring,
- Kaytlyn A. Gerbin,
- Tanya Grancharova,
- Benjamin W. Gregor,
- Lisa J. Harrylock,
- Amanda Haupt,
- Melissa C. Hendershott,
- Caroline Hookway,
- Alan R. Horwitz,
- H. Christopher Hughes,

- Eric J. Isaac,
- Gregory R. Johnson,
- Brian Kim,
- Andrew N. Leonard,
- Winnie W. Leung,
- Jordan J. Lucas,
- Susan A. Ludmann,
- Blair M. Lyons,
- Haseeb Malik,
- Ryan McGregor,
- Gabe E. Medrash,
- Sean L. Meharry,
- Kevin Mitcham,
- Irina A. Mueller,
- Timothy L. Murphy-Stevens,
- Aditya Nath,
- Angelique M. Nelson,
- Sandra A. Oluoch,
- Luana Paleologu,
- T. Alexander Popiel,
- Megan M. Riel-Mehan,
- Brock Roberts,
- Lisa M. Schaefbauer,
- Magdalena Schwarzl,
- Jamie Sherman,
- Sylvain Slaton,
- M. Filip Sluzewski,
- Jacqueline E. Smith,
- Youngmee Sul,
- Madison J. Swain-Bowden,
- W. Joyce Tang,
- Derek J. Thirstrup,
- Daniel M. Toloudis,
- Andrew P. Tucker,
- Veronica Valencia,
- Winfried Wiegraebe,
- Thushara Wijeratna,

- Ruian Yang,
- Rebecca J. Zaunbrecher,
- Ramon Lorenzo D. Labitigan,
- Adrian L. Sanborn,
- Graham T. Johnson,
- Ruwanthi N. Gunawardane,
- Nathalie Gaudreault,
- Julie A. Theriot &
- …
- Susanne M. Rafelski

Show authors

Nature **volume 613**, pages 345–354 (2023)

- 13k Accesses
- 418 Altmetric
- Metrics details

Abstract

Understanding how a subset of expressed genes dictates cellular phenotype is a considerable challenge owing to the large numbers of molecules involved, their combinatorics and the plethora of cellular behaviours that they determine[1,2]. Here we reduced this complexity by focusing on cellular organization—a key readout and driver of cell behaviour[3,4]—at the level of major cellular structures that represent distinct organelles and functional machines, and generated the WTC-11 hiPSC Single-Cell Image Dataset v1, which contains more than 200,000 live cells in 3D, spanning 25 key cellular structures. The scale and quality of this dataset permitted the creation of a generalizable analysis framework to convert raw image data of cells and their structures into dimensionally reduced, quantitative measurements that can be interpreted by humans, and to facilitate data exploration. This framework embraces the vast cell-to-cell variability that is observed within a normal population, facilitates the integration of cell-by-cell structural data

and allows quantitative analyses of distinct, separable aspects of organization within and across different cell populations. We found that the integrated intracellular organization of interphase cells was robust to the wide range of variation in cell shape in the population; that the average locations of some structures became polarized in cells at the edges of colonies while maintaining the ‘wiring’ of their interactions with other structures; and that, by contrast, changes in the location of structures during early mitotic reorganization were accompanied by changes in their wiring.

Main

Cellular organization can be defined as the sum total of how all of a cell’s components are arranged within it, generating an overall characteristic size, shape and appearance for a cell of a given type. The models and laws for understanding and predicting cellular organization and its pivotal role as a determinant of cellular phenotype remain to be determined. A first step towards this goal is to identify interpretable and testable principles, or ‘rules’, that govern cell organization. One approach is through a systematic analysis of the locations and quantitative relationships among many different cellular structures within large populations of cells and how these relationships vary with the morphology and behaviour of the cell itself. To define cell organization precisely and quantitatively, however, requires measuring multiple distinct aspects of organization; for example, the size (or number) and shape of each structure, its locations in the cell, its direct and indirect interactions with all the other structures and the temporal changes. A population of putatively identical cells might exhibit substantial cell-to-cell variability as they respond sensitively to their ever-changing internal and external contexts, such as the cell cycle, differentiation or changes in their environment. Furthermore, an abnormal cell quantitative phenotype might exhibit not only a shift in the mean but also a shift in the variability[5]. Thus, a meaningful description of cell organization requires a formal definition and categorization that includes robust, objective and quantitative measurements of both the mean and the variability in the descriptors of organization.

Creating such a nuanced, formal quantitative view of cell organization will enable the statistical comparisons that identify generalizations and elucidate how cell organization differs within and across different cell populations and

during transitions among normal or abnormal cell behaviours. It will also permit deeper investigations that integrate cell organization with cell behaviour and cell identity, including the integration of distinct data types (for example, images and various ‘omics), leading to more meaningful and useful definitions of cell types and states[6,7,8,9,10,11,12].

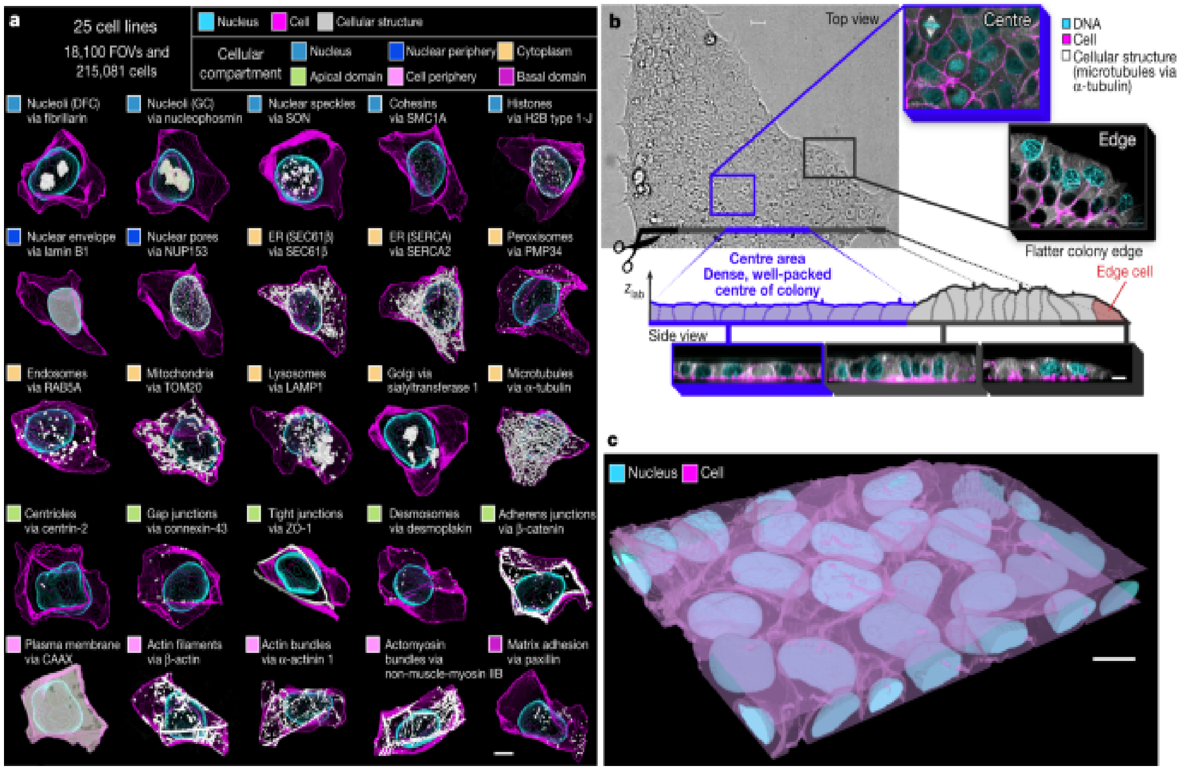
We have initiated our study of cellular organization by focusing on the integrated organization of 25 cellular structures that represent major intracellular machines and organelles, generating an extensive, high-replicate baseline dataset of 3D live-cell images of normal human induced pluripotent stem cells (hiPS cells): the WTC-11 hiPSC Single-Cell Image Dataset v1. This dataset was used to develop a generalizable and extensible quantitative analysis framework based on two conceptually distinct coordinate systems to analyse the cells. The first coordinate system defines the cell and nuclear shape of each individual cell with respect to the total variation in the observed population. The second coordinate system specifies the spatial location of every cellular structure within an individual cell. When combined, the two coordinate systems permitted the development of a suite of statistical measurements to quantify distinct aspects of cell organization, formally distinguishing among three kinds of change while controlling for the effects of natural cell-shape variation: (1) changes in the average location of individual structures; (2) changes in the variability of these locations; and (3) changes in the pairwise interactions among structures. We applied our framework to three subsets of cells in the dataset—the large baseline population of cells in interphase, the cells at the outer edges of the epithelial-like hiPS cell colonies, and cells undergoing reorganization during early mitosis—and developed data-visualization approaches to summarize these results in a way conducive to data exploration.

WTC-11 hiPSC Single-Cell Image Dataset v1

hiPS cells represent an early embryonic cell state and are a useful model system for human cells. hiPS cells are naturally immortal, karyotypically normal and can be induced to differentiate into other cell types[13]. We previously developed methods and quality-control workflows to create the ten inaugural hiPS cell lines in the Allen Cell Collection[14] (described at

https://www.allencell.org/cell-catalog.html), each expressing a single endogenously tagged protein representing a particular organelle or cellular structure. For this study, we created 15 new Allen Cell Collection lines that provide a holistic view of cells at the level of 25 of their major organelles, cellular structures and compartments (Fig. 1a). We built an automated and standardized microscopy imaging pipeline to generate the living colonies, imaged the cells in 3D using spinning-disk confocal microscopes and then processed the images to create the WTC-11 hiPSC Single-Cell Image Dataset v1 (Fig. 1 and Extended Data Fig. 1). We included fluorescent cell-membrane and DNA dyes to reference the locations of fluorescent protein (FP)-tagged cellular structures relative to the cell boundary and the nucleus or mitotic chromosomes. For each of the 25 cellular structures, we used 3D segmentations of the tagged protein to identify the location and morphology of the structure itself, rather than the location of the FP-tagged protein signal (Extended Data Fig. 2). The tightly packed, epithelial-like nature of hiPS cells, as well as the need for highly accurate 3D cell boundaries to minimize the misassignment of cellular structures to neighbouring cells required deep-learning-based segmentation approaches to create a robust, scalable and highly accurate 3D cell and nuclear segmentation algorithm[15] (Methods), which was applied to all 18,100 fields of view (FOVs) to extract the 215,081 cells presented in this dataset (Extended Data Fig. 1). Both the FOV images and the single-cell dataset are available as downloadable files (see Data availability) and through interactive online visual-analysis tools that require no software installation or expertise (https://cfe.allencell.org/). For the analyses described below, we used subsets of the dataset including the baseline interphase dataset (202,847 cells), cells at the edge of colonies (5,169 cells) and cells in early mitosis (3,182 cells) (Extended Data Fig. 1d).

Fig. 1: The WTC-11 hiPSC Single-Cell Image Dataset v1 includes 25 cell lines that represent key cellular structures located throughout all of the major compartments of the cell.

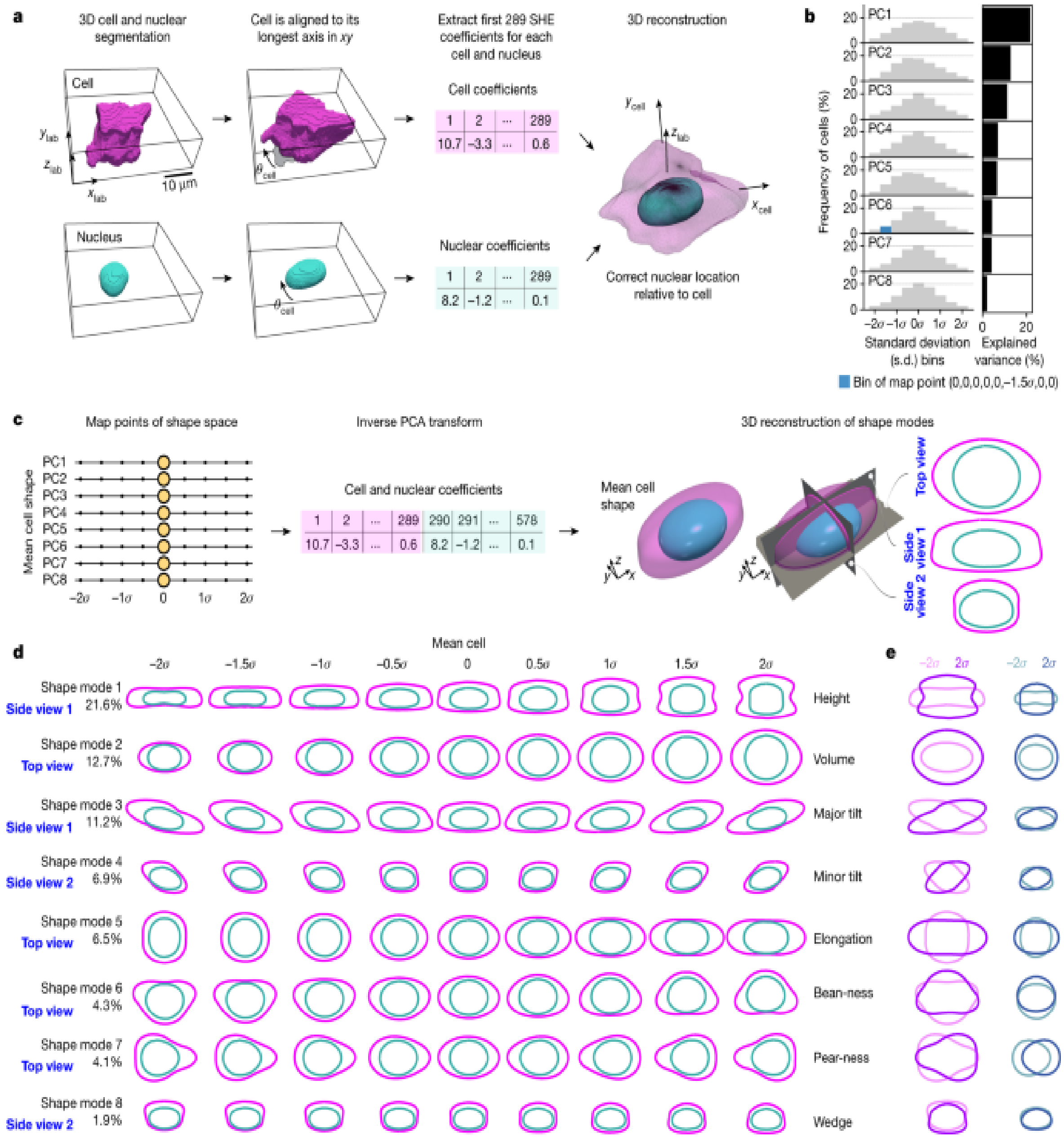


a, Maximum intensity projections of one representative cell example per cellular structure, based on segmentations of the structure (white), the cell membrane (magenta) and the DNA (cyan). The fluorescently tagged protein representing the structure and the cellular compartment (Fig. 3d) are indicated. DFC, dense fibrillar component; ER, endoplasmic reticulum; GC, granular component. **b**, Top and side views (single slice) of hiPS cells with FP-tagged microtubules (via α-tubulin), grown in tightly packed, epithelial-like colonies and labelled with cell-membrane (magenta) and DNA (cyan) dyes to permit imaging and segmenting of cells and nuclei. Cells were most frequently imaged halfway towards the centres of large, well-packed colonies (blue) where they behave most consistently, but were also imaged at other locations within the colony, such as at the edges of colonies (red). z_{lab} denotes the lab frame of reference. **c**, Three-dimensional visualization of cell and DNA segmentations within a colony of hiPS cells. Total numbers of acquisition days, FOVs and cells per cellular structure are in Supplementary Data 1 and Extended Data Fig. 1d. Scale bars, 10 μm.

A PCA-based cell and nuclear shape space

To embrace the great diversity of the 202,847 3D images of cells in interphase spanning 25 cellular structures, and to directly compare cellular organization across this large population, we built a cell and nuclear shape-based coordinate system (Fig. 2), adapting a standard principal component analysis (PCA)-based dimensional reduction approach[16]. We aligned all cells along their longest axis in the *xy* plane, preserving their biologically relevant, epithelial-like apical–basal axis. We then used a spherical harmonic expansion (SHE)[17,18] to accurately parameterize each 3D cell and nuclear shape with a set of orthogonal periodic basis set functions, defined on the surface of a sphere (Fig. 2a and Extended Data Fig. 3). The joint vectors for all cells (578 SHE coefficients) were then subjected to PCA. We found that the first eight principal components represented about 70% of the total variance in cell and nuclear shape (Fig. 2b). Thus, with this dimensionality reduction, the cell and nuclear shapes for each individual cell can be approximately reconstructed from a small vector with only eight components. This dimensionality reduction also organizes the cells into a simple, intuitive eight-dimensional (8D) generative 'shape space'. For example, the origin (0,0,0,0,0,0,0,0) of the shape space can be reconstructed through the values of the SHE coefficients representing this location in the 8D coordinate system, and can then be visualized as an idealized cell shape that statistically represents the average, or mean, shape ('mean cell shape') of all of the cells in the dataset (Fig. 2c). Similarly, idealized shapes can be reconstructed by traversing across each of the eight orthogonal axes in the shape space (Extended Data Fig. 3b).

Fig. 2: A PCA-based cell and nuclear shape space reveals interpretable modes of hiPS cell-shape variation.



a, Segmented 3D images of a cell and its nucleus are rotated in the *xy* plane by θ_{cell} degrees around the cell centroid such that the longest axis of the cell is parallel to the *x* axis. These aligned images are the input for SHE of degree $L_{max} = 16$, resulting in a total of 578 SHE coefficients (289 for each the cell and the nucleus), which are used to reconstruct the cell and nuclear shape and nuclear location with high accuracy. x_{lab}, y_{lab} and z_{lab} denote the lab frame of reference and x_{cell} and y_{cell} the rotated cell frame of reference. Scale bar, 10 μm. **b**, Frequency of cells per map point bin (left) and

explained variance (right) for the first eight principal components (PCs) of the PCA applied to the SHE coefficients for interphase cells ($n = 202,847$). Blue denotes one map point bin. **c**, Eight shape modes comprise the cell and nuclear shape space. Each is a normalized PC (standard deviation (s.d.), σ, units) sampled at nine map points (-2σ to 2σ in steps of 0.5σ). Three-dimensional shape reconstructions can be created at each of these map points—here yellow dots at the origin (0,0,0,0,0,0,0,0)—using an inverse PCA transform and its resultant SHE coefficients. Three 2D views of the 3D shape are shown. **d**, Most relevant 2D view of 3D shapes reconstructed at each of the nine map points for each of the eight shape modes (given names that can be interpreted by humans). Supplementary Video 1 shows all three 2D views. The centre bin in all modes is the identical mean cell shape. **e**, Overlay of 2D views of the cell (magenta) and nucleus (cyan) for the two most extreme map points (at -2σ, lighter, and 2σ, darker) of each shape mode.

Source data

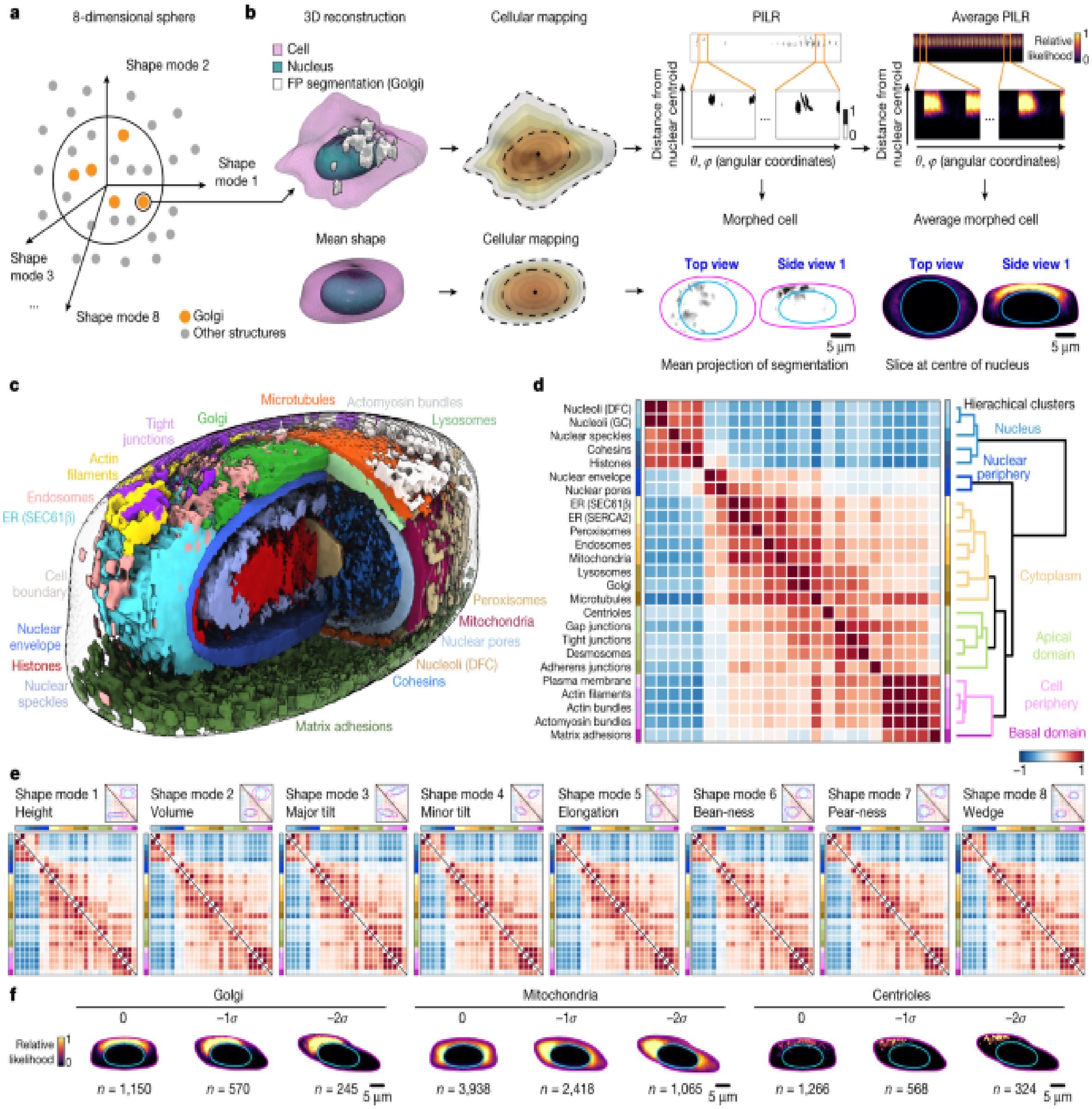
To build a human-interpretable understanding of the modes of shape variation in our population, we reconstructed cell and nuclear shapes at regular intervals along every axis of this shape space (Fig. 2d and Supplementary Video 1). These idealized cells represent ‘map points’ within the shape space that can be used to identify and cluster individual real cells that are similar in shape to each idealized map point and to each other. Intuitively, these mathematically orthogonal modes of shape variation appear to describe expected variable cell-shape features that are independent of one another. Shape mode 1 appeared to largely reflect the height of the cell (Extended Data Fig. 3c), which was mainly determined by the surface area of the hiPS cell colony and the position of the cell in the colony (Supplementary Methods). Shape mode 2 appeared to largely reflect the overall volume of the cell, representative of cell-cycle progression. The correlation between cell height and cell volume was relatively modest ($R = 0.34$; Extended Data Fig. 3c), meaning that cells with a given height may have a wide range of volumes and vice versa. Shape mode 1 and shape mode 2 thus disentangle cell volume and cell height from each other. The remaining shape modes 3 to 8 represented other systematic ways in which the shapes of these epithelial-like cells might vary, such as tilting along the major or minor *xy* axes (shape modes 3 and 4). In shape modes 1, 2 and 5,

nuclear shape changed concomitant with cell shape, whereas for the other shape modes it was the position and orientation of the nucleus within the cell that adjusted concomitant with cell shape (Fig. 2d,e and Supplementary Video 1).

Integrated average morphed cells

This standardized cell and nuclear shape space permits the clustering of similarly shaped cells and thus facilitates an investigation of the location of cellular structures within the confines of cells with similar 3D outer cell boundaries and nuclear spatial constraints. For example, to determine the average locations of cellular structures within the mean cell and nuclear shape, we first identified all of the cells within an ‘8-dimensional sphere’ with its origin at the very centre of the shape space encompassing the 35,636 cells that lie in this region closest to this origin (Fig. 3a and Methods). To directly and quantitatively compare the locations of each of the 25 cellular structures in these relatively similarly shaped individual cells, we developed an intracellular location coordinate system that took advantage of the SHE describing the outer cell boundary, the outer nuclear boundary and the centre of the nucleus and then interpolated between the relevant SHE coefficients. This permitted us to map the presence or absence of a structure within an individual cell at all of the possible points along these concentric 3D shells and store this information in a parameterized intracellular location representation (PILR). We could then ‘morph’ the locations of this structure, through the PILR, into the equivalent locations in an identically bounded reconstructed cell shape that represents that cell’s actual shape (Fig. 3b, Extended Data Fig. 4 and Methods). For each structure, we averaged the PILRs across all of the similarly shaped cells in the 8-dimensional sphere and then morphed the average PILR into the equivalent locations within the mean cell shape, creating the ‘average morphed cell’ for that structure (Fig. 3b and Extended Data Fig. 5). These average morphed cells represent the relative likelihood of a structure being at a location in the cell, conceptually similar to previous approaches that have been used to analyse images of cells grown on micropatterns[19]. We then combined these 25 average morphed cells to create an integrated visualization of the average locations of all 25 structures (Fig. 3c and Supplementary Video 2).

Fig. 3: Creating an average pairwise spatial interaction map of cellular structures.



a, Diagram illustrating the clustering of the 35,633 cells closest to the origin of an 8-dimensional sphere centred at the origin of the shape space. **b**, Creating average morphed cells. Top left, 3D visualization of the segmentations of a cell (magenta), nucleus (cyan), and cellular structure (here Golgi in white). Bottom left, the equivalent for the mean cell and nuclear shape. ‘Cellular mapping’ shows the results of interpolating the SHE coefficients to generate successive 3D concentric mesh shells (different colours) from the centroid of the nucleus (black dot) to the nuclear (inner)

and then to the cell (outer) boundary to create the nuclear and cytoplasmic mapping, respectively. The presence or absence of the structure is recorded at each mesh point location, resulting in a PILR, shown in matrix format for the Golgi of this cell. The PILR of an individual cell or the 'average PILR' of the 1,058 Golgi-tagged cells within the 8-dimensional sphere can be mapped into the mean cell and nuclear shape, generating 'morphed' and 'average morphed' cells, respectively. Scale bars, 5 µm. **c**, Integrated 3D visualization of 17 of the 25 structures to illustrate their average relative spatial relationships (Supplementary Video 2). **d**, Average pairwise spatial interaction map of cellular structures. Heat map of the average location similarities (Pearson correlations between average PILRs; Extended Data Fig. 4g) for every pair of 25 cellular structures for cells in the 8-dimensional sphere. A clustering algorithm generates the dendrogram (left) with coloured branches of the six top-level clusters lengths representing the distance between clusters. **e**, Average spatial interactions are robust to systematic variations in cell and nuclear shape. Heat maps for the -2σ (bottom triangle) and 2σ (top triangle) shape space map points for each of the eight shape modes (numbers of cells and heat map data in Supplementary Data 1). **f**, Side view 1 of average morphed cells for three structures and three bins (0, -1 and -2σ) along shape mode 3 (major tilt). Scale bars, 5 µm.

Source data

Average pairwise spatial interaction map

To measure the relationships of the average locations of each of the 25 cellular structures relative to all the others after computational integration, we calculated the 2D pixel-wise Pearson correlation between the averaged PILRs for all pairs of structures within the 8-dimensional sphere, representing a measure of the 'average location similarity' between two structures (Extended Data Fig. 4g). In principle, the overall average location similarity among structures could span a range. At one extreme, all structures could be coupled, for example, every structure depending on every other structure, whereas at the other extreme, every structure could be independent from every other structure. We performed a hierarchical clustering analysis of these correlation values to create a purely data-driven 'average pairwise spatial interaction map' of cellular structures. Notably, we

found that the cellular structures clustered naturally into an ordered radial compartmentalization of the cell, from the centre of the nucleus outward (Fig. 3d), and also separated between the apical and basal domains of the cell. The six top-level clusters included structures localized to the nucleus, nuclear periphery, cytoplasm, apical domain (in a dispersed way), cell periphery and basal domain, respectively. The spatial interaction map hierarchy confirmed the expected strong location similarities within several sets of cellular structures (for example, two nucleolar structures (DFC and GC), two ER structures (SEC61β and SERCA) and three actin-related structures (actin filaments, actin bundles and actomyosin bundles)), validating this analysis approach. We found a very high location similarity between lysosomes and Golgi, consistent with their enrichment in location in the apical cytoplasm and the known role of the Golgi in regulating lysosome localization[20,21]. Mitochondria shared greater location similarity with the ER (SEC61β and SERCA) than any other structures, consistent with the functional interactions between these cellular structures[22,23].

Average spatial interactions are robust

The ability to analyse the location of cellular structures throughout a human-interpretable standardized cell and nuclear shape space allows us to ask how robust the relative average locations of cellular structures are when they are subjected to the systematic variation in cell and nuclear shape that is present in this dataset. For example, we can compare differences in average structure locations between flat and tall cells, small and large cells or cells with shapes that are less or more polarized. We clustered all cells in the dataset into 9 bins along each of the 8 shape modes in regular intervals (as in Fig. 2) to create a total of 65 cell-shape map points (the centre bin is the same in all modes), into which we morphed each of the 25 structures (Supplementary Video 3). Of note, we found very little change in the overall average location interaction map of these 25 structures throughout the shape space (Fig. 3e and Extended Data Fig. 4h). Instead, structures filled whatever cytoplasmic space was available to them in the particular shape while maintaining their appropriate apical–basal localization and their relative average locations (three examples for shape mode 3 in Fig. 3f; all 25 structures through the shape space in Supplementary Video 3).

Variations in structure locations

The combination of the two coordinate systems—the shape space and the PILR—creates an analysis framework to investigate not only the average locations and pairwise interactions, but also their variability. We calculated the 2D pixel-wise Pearson correlation between the PILRs for all pairs of the 35,636 individual cells, including all 25 cellular structures, within the 8-dimensional sphere centred at the mean cell and nuclear shape, regardless of whether any 2 cells have the same or different tagged structures. This creates a matrix of pairwise structure PILR correlation values for all pairs of individual cells (Extended Data Fig. 6a). Correlation values from this matrix can then be averaged within all pairs of structures to create an average correlation matrix to obtain two distinct measurements of structure location and its variability: the 'location stereotypy' and the 'location concordance' (Extended Data Fig. 6b). The diagonal of this matrix is the location stereotypy; that is, the average of all the pairwise PILR correlation values for a given structure. Structures with a high stereotypy value have little cell-to-cell variability in their overall absolute positions, whereas structures with a low stereotypy value may be more often found in distinct locations amongst different cells. Comparing the stereotypy for each structure permitted us to rank structures that are most to least stereotyped in their locations within the mean cell and nuclear shape (Extended Data Figs. 6b and 7a).

The off-diagonal values in the average correlation matrix are the location concordances between pairs of structures—a measure analogous to the stereotypy, but representing aspects both of how similar the absolute locations of two structures are and how variable those relative locations may be among different cells (Extended Data Fig. 6b). For example, in the average spatial interaction map (Fig. 3d), the average location of peroxisomes was more similar to that of other cytoplasmic organelles (endosomes and mitochondria) than structures in the nucleus or at the cell periphery, and this relationship is maintained in the concordance between these structures. However, from cell to cell, the absolute locations of peroxisomes are very variable (amongst the lowest stereotypy owing to their sparse, punctate nature) and thus their concordance with other cytoplasmic structures (that is, the correlation between their absolute locations) is very

low. We investigated how much the stereotypy and concordance changed in response to changes in cell shape and found that in addition to the average pairwise structure locations (Fig. 3e), the variability in individual and pairwise structure locations was also extremely robust to overall cell-shape variation in this dataset (Extended Data Figs. 6c,d and 7, Supplementary Data 1 and Methods).

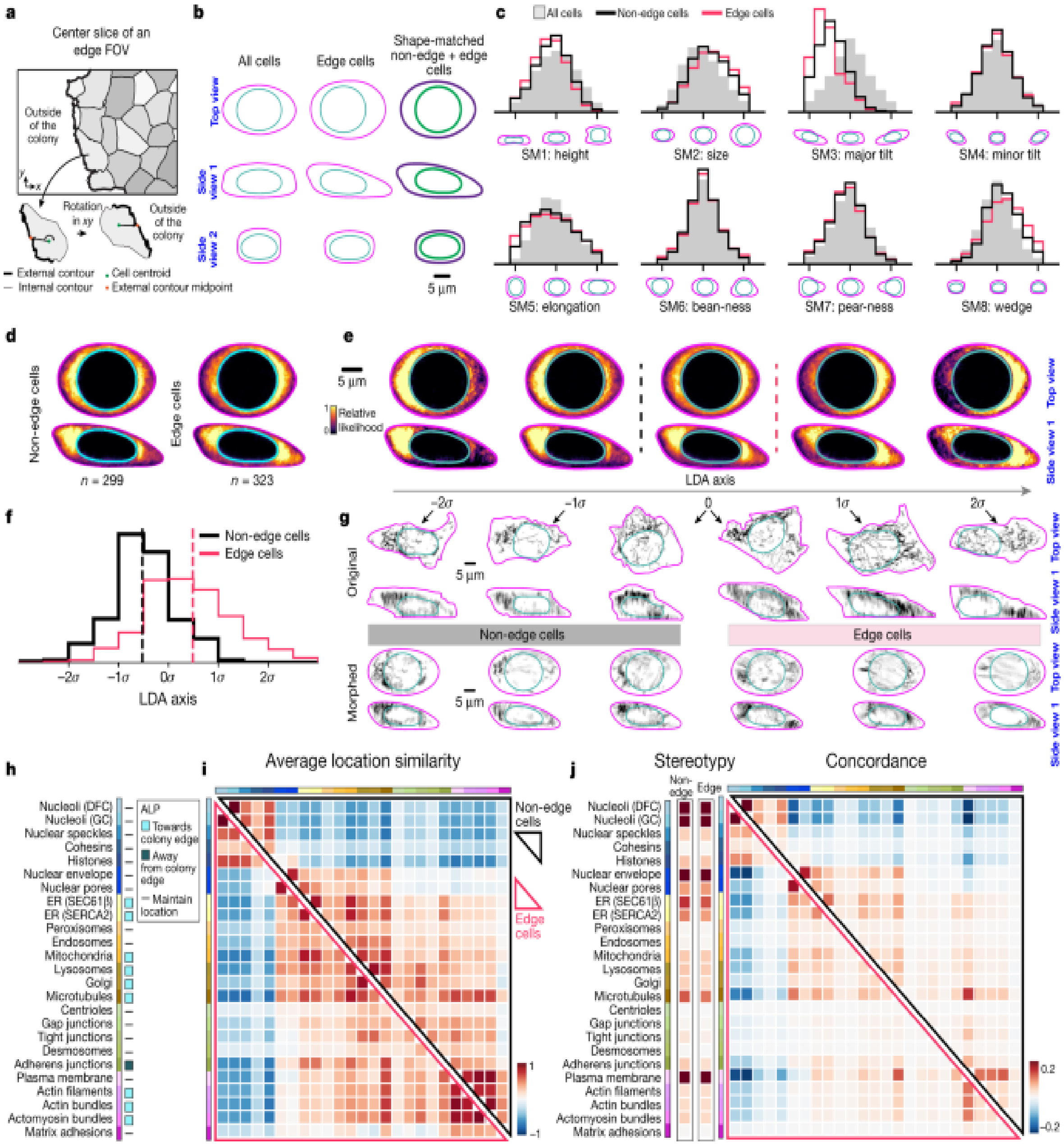
Systematic analysis of structure size

Cellular structures also exhibit cell-to-cell variability in their structure size (or number). It has previously been shown that the volume of several cellular structures in the cell correlates with the overall cell volume, including the nucleus and mitochondria[24]. We therefore used our large dataset to perform a systematic and comparative analysis of the relationship between cellular structure volume and five relevant size metrics (cell volume, cell surface area, nuclear volume, nuclear surface area and cytoplasmic volume) for 15 of the cellular structures in this dataset (Extended Data Fig. 8). Although nuclear structures seemed to be most tightly coupled to nuclear size metrics, cytoplasmic structures ranged more widely in how well the variance in their volumes was uniquely attributable to cell versus nuclear size metrics. Unexpectedly, the variance in nuclear speckle (SON) volumes was most uniquely attributable to the nuclear surface area and not the nuclear volume, although speckles localize throughout the nucleoplasm. This is notable in light of the possible connection between transcript splicing (which occurs at nuclear speckles) and increased rates of nuclear export[25]. We found that contributions from other shape modes were negligible (Extended Data Fig. 8), suggesting that cell and nuclear size, and not other aspects of shape, affect the variability in the size of cellular structures. Overall, these results show that the degree to which cell and nuclear size metrics account for the variation in cytoplasmic structure volumes is structure dependent, consistent with the wide range of cell functions that these structures regulate.

Polarized reorganization in edge cells

Most cells within the tightly packed, epithelial-like hiPS cell colonies form cell–cell contacts with their neighbouring cells in a continuous circumferential band. Cells located at the edges of colonies (edge cells), however, have a distinct morphology because they lack cell–cell junctions along their outermost edge and have been shown to differ in their transcriptional profiles and metabolic activity[26,27]. To determine whether, and precisely how, the cellular organization of edge cells differs from that of cells not at the edge, we extended the two-coordinate-system analysis framework to permit the comparative analysis of integrated cellular organization in a second, distinct cell population within the dataset. We aligned edge cells such that their positive *x* axis was oriented towards the outer edge of the colony (Fig. 4a), and then mapped them into the baseline cell and nuclear shape space. On average, consistent with expectations, edge cells were much more tilted than the baseline interphase population (Fig. 4b,c). To directly compare cellular organization in similarly shaped cells, we took advantage of the very large size of the baseline dataset to identify a set of non-edge cells that were the most similarly shaped to each edge cell (Extended Data Fig. 9a). The resultant 'shape-matched' dataset comprises two distinct populations—edge and non-edge cells—with almost identical cell-shape distributions (Fig. 4b,c).

Fig. 4: Cellular structure locations are polarized but cellular structure location wiring is unaltered in cells at the edge of hiPS cell colonies.



a, Alignment. Cells at the edge of the colony are rotated in *xy* so that the axis between the cell centroid and the external contour midpoint is parallel to the *x* axis and the outer contour edge of the cell is oriented to the right. **b**, Mean cell (magenta or purple) and nuclear (cyan or green) shape for all interphase cells (left), edge cells (centre) and the shape-matched non-edge cells and edge cells combined. Three 2D views of the 3D shape are shown. Scale bar, 5 µm. **c**, Frequency of cells for the eight shape modes (SM) for all interphase (grey), non-edge (black) and edge (red) cells. **d**, Average morphed cells for mitochondria in non-edge and edge cells. **e**, ALP via LDA. PILR-LDA-

based reconstructions of mitochondria in average morphed cells at five positions (in σ units) along the LDA axis. Dotted lines correspond to the locations of the mean non-edge (black) and edge (red) cells in **d**. **f**, Frequency of cells along the LDA axis within non-edge and edge cell populations. Dotted vertical lines indicate the means. **g**, Top view and side view 1 of three examples of non-edge and edge cells along the LDA axis. Top row shows the original and bottom row the morphed visualizations for each of these cells. Images are average projections of the segmented structure. **h**, The ALP for 25 cellular structures in edge cells. **i**,**j**, Heat maps of the average location similarity (**i**), stereotypy (**j**, left) and concordance (**j**, right) in non-edge cells (top triangle or left column in stereotypy) and edge cells (bottom triangle or right column in stereotypy). Numbers of cells and heat map data are in Supplementary Data 1. Scale bars, 5 μm.

Source data

We compared the average locations (through the average morphed cells) of the 25 structures in edge cells and shape-matched non-edge cells (Fig. 4d, Extended Data Fig. 9b,d and Supplementary Video 4). We found a noticeable polarization of cytoplasmic structures and organelles (for example, mitochondria, microtubules, lysosomes and Golgi apparatus), as well as structures representing the actin cytoskeleton (for example, actin filaments, actin and actomyosin bundles) towards the outer periphery of edge cells. Adherens junctions were polarized away from the colony periphery, supporting the lack of cell–cell junctions at the edge. To quantify the changes in the locations of cellular structures between the two distinct shape-matched populations, we took advantage of the PILR as a high-dimensional representation of the intracellular location space. We reduced the PILR dimensionality down to the primary axis of greatest difference in intracellular location for each cellular structure, first through a PCA and then by a linear discriminant analysis (LDA) to identify the linear combination of PCs that best separates non-edge and edge cells (Extended Data Fig. 9c and Methods). We then reconstructed PILRs and generated morphed cells at positions along the one-dimensional LDA axis representing the full range of the location phenotype for each structure. For example, the more versus less polarized location phenotypes of mitochondria and actin bundles seen in the average morphed cells could be reconstructed at their appropriate positions along the LDA axis and the polarized nature of this location phenotype

extrapolated by comparing the reconstructions further away from the means (Fig. 4d,e, Extended Data Fig. 9d,e and Supplementary Video 4). Individual cells could now also be sorted along this LDA axis and further analysed, for example through histograms that represent the entire edge and non-edge cell populations (Fig. 4f,g and Extended Data Fig. 9f–h). The PILR-LDA approach, together with visual assessment of average and individual morphed cells (Methods), permitted the determination of the biological average location phenotype (ALP) for each of the 25 cellular structures in edge cells (Fig. 4h). This analysis confirmed a polarized relocation of cytoplasmic organelles and actin cytoskeletal structures towards the edges of colonies in edge cells when compared with shape-matched non-edge cells. Thus, cell shape alone does not drive integrated intracellular organization.

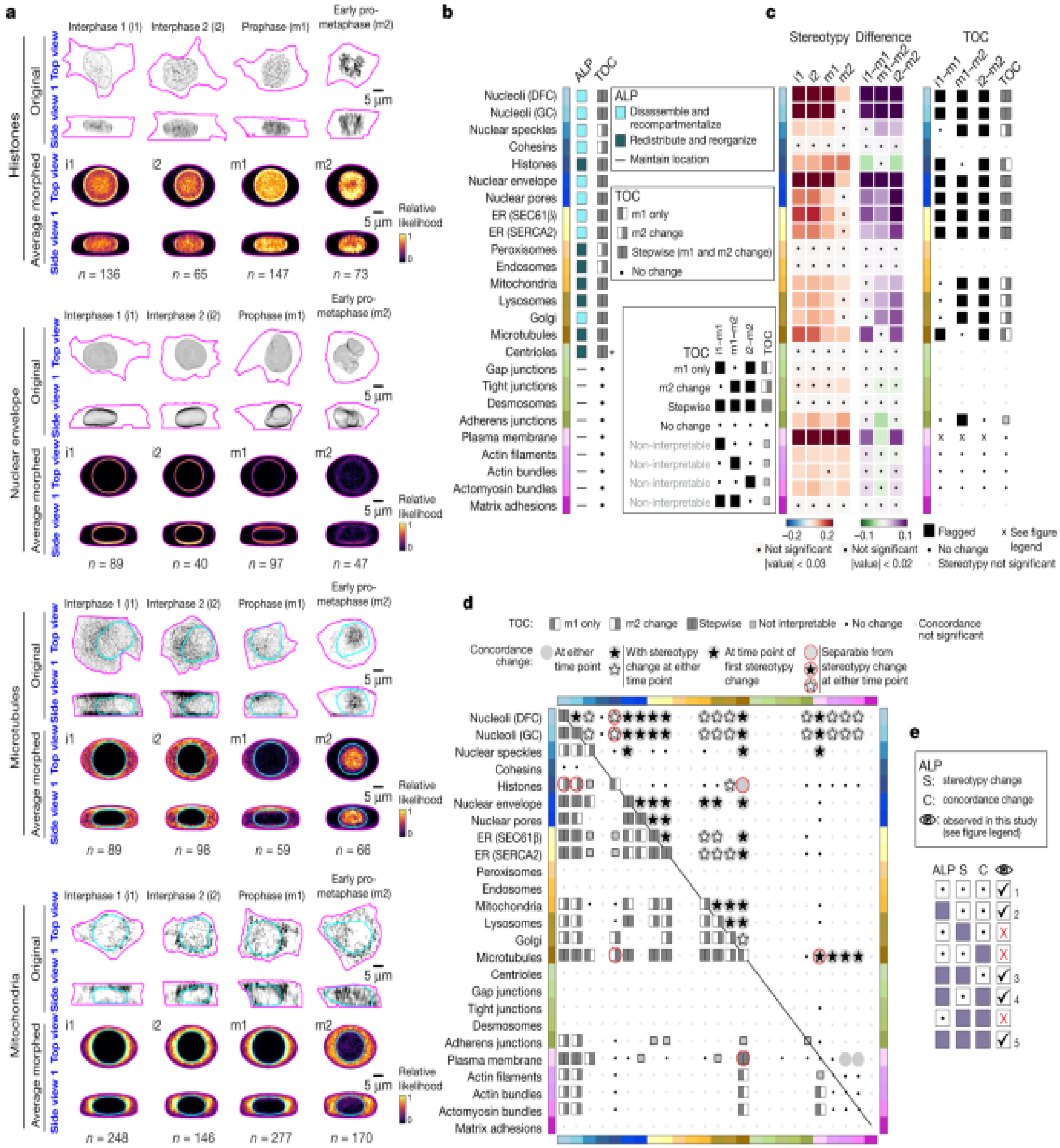
We compared the average location similarities, stereotypy and concordance between edge and non-edge cells and found little—if any—differences in these (Fig. 4i,j and Extended Data Fig. 9i), despite the ALPs found in edge cells for many of these structures. We also compared the average structure volumes (15 structures validated for volume analysis) and found very few changes between edge and non-edge cells (Methods and Supplementary Data 1). One notable result, however, was that the median volume ratio of mitochondria relative to cell size was greater in edge cells (0.099; $n = 322$) than in non-edge cells (0.087, $n = 299$; 14% effect size increase, rank-sum test $P = 9.2 \times 10^{-11}$). These results may reflect previous observations of differences in mitochondrial protein composition and function in colony edge cells[26] and of mitochondrial abundance in cells grown at different densities[9]. Overall, these results suggest that although the average location of many cellular structures is changed in edge cells, the relative wiring of these structures to each other and the extent to which their locations vary is maintained. This suite of measurements thus facilitates a more nuanced identification of which distinct aspects of integrated intracellular organization are changed between different populations of cells, instead of a more generic change in cellular organization.

Integrated early mitotic reorganization

We took advantage of the marked intracellular reorganization that occurs as cells enter mitosis[28] to further examine the relationship between our suite of

measurements of the average and relative locations and the variability of cellular structures. We focused on the two earliest stages of mitosis—prophase (m1) and early prometaphase (m2), when the condensing chromosomes still largely form an aggregated, nuclear-like structure that could be biologically interpreted in the context of our interphase (i) cell and nuclear shape-based coordinate system (Extended Data Fig. 10a). We mapped the shapes of m1 and m2 cells into the cell and nuclear shape space. Although cells in m1 were generally larger than average interphase cells, as expected, they were also mostly of similar overall shape to cells in interphase. By m2, however, cells exhibited mitosis-related changes in shape, including increased height and a more uniform rounder cell shape. Analogously to our analysis of edge cells, we created the appropriate shape-matched datasets for m1 and m2, matched to interphase cell subsets i1 and i2, respectively (Extended Data Fig. 10b,c). We extended the analysis framework to incorporate a time component through four timing of change (TOC) categories (Fig. 5b–d and Methods) permitting the analysis of intracellular reorganization over three sequential cell-cycle stages. We also developed a standardized process to systematically identify and flag all entries in the average correlation matrix (stereotypy and concordance values) that changed in a significant way between two conditions (Extended Data Fig. 10d–f and Methods). This approach permitted us to determine whether and when structures underwent a change in their individual or pairwise relative locations or in the variability in these locations.

Fig. 5: Integrated intracellular reorganization in early mitosis.



a, Individual cell examples (top) and average morphed cells (bottom) for four cellular structures in prophase (m1) and early prometaphase (m2), shape matched to interphase cell subsets i1 and i2, respectively (Extended Data Fig. 10b). Cyan DNA outlines were left out for the histones and nuclear envelope to better see their locations at the nuclear periphery. Scale bars, 5 µm. **b**, The ALP and its timing of change (TOC) for 25 cellular structures in early mitosis. Asterisk indicates centriole ALP determined by visual inspection (Supplementary Methods). **c**, Left, heat maps of stereotypy (blue to red) and stereotypy differences (green to purple) in early mitosis.

Black dots indicate values below the measurable cut-offs (Methods). Right, flagged significant stereotypy differences for each structure between interphase and both early mitotic stages (filled black boxes) as well as the resultant stereotypy TOC. The stereotypy of the plasma membrane was so high that, although the absolute difference in stereotypy values passed the flag criteria, the relative values were extremely small (denoted with 'x'). **d**, Timing and types of change in concordance, through the PILR average correlation matrix. Bottom triangle: the concordance TOC assignments for all pairs of structures. Heat maps of intermediate steps are in Extended Data Fig. 10d–g and Supplementary Data 1. Top triangle: types of changes in concordance relative to changes in stereotypy as described in the results (Methods). Numbers range from $n = 6$ to 256 cells depending on the structure and stage (Supplementary Data 1). Coloured bars at the left of heat or colour maps in **b**–**d** indicate the cellular structure. Owing to the low number of cells in mitosis for some structures, we could not quantitatively analyse differences in the average location similarities, although their qualitative results matched those based on the concordance values (Extended Data Fig. 10g). **e**, Summary of examples of changes in distinct aspects of organization observed throughout this study. Specific examples are indicated with numbers: (1) structures that maintained locations in edge cells and early mitosis; (2) structures that polarized in edge cells; (3) for example, histones and microtubules at m1; (4) for example, histones and microtubules at m2; (5) most structures during early mitosis.

Source data

We found that the ALPs of the 25 structures fell into three classes (also at https://imsc.allencell.org/): (1) the locations of structures at the cell periphery for example, the plasma membrane, actin-related structures, cell–cell adhesions) were largely maintained; (2) most structures within or surrounding the nucleus (for example, nucleoli, nuclear envelope and ER) disassembled and the FP-tagged proteins were recompartmentalized; and (3) most structures within the bulk of the cytoplasm (for example, mitochondria and lysosomes) reorganized and redistributed throughout the cytoplasm as the microtubules themselves reorganized and redistributed towards the condensing chromosomes and the centre of the cell (Fig. 5a,b and Supplemenrary Video 5). Almost all structures that changed locations in early mitosis did so both in m1 and in m2 (stepwise TOC category).

Exceptions included nuclear speckles and cohesins, which did not change location until m2, when they began to disassemble. This was at a later stage of mitosis than all of the other nuclear and nuclear periphery structures. Peroxisomes and endosomes also did not noticeably redistribute towards the centre of the cell until m2.

For most nuclear and nuclear periphery-related structures (for example, nuclear envelope, speckles and ER), a change in their location coincided with a change in how variable that location was (for example, a matched TOC for ALP and stereotypy in Fig. 5b,c). However, for other structures, including most of the cytoplasmic structures, there was a discrepancy between the timing of change in average location and its variability (for example, mitochondria, Golgi, lysosomes in m1 as well as histones and microtubules in m2). Some structures that did not show any changes in stereotypy were discrete, punctate structures with very low stereotypy, for which changes in stereotypy could not be determined with statistical confidence (for example, cohesins, endosomes and desmosomes). All of the structures that maintained their location at the cell periphery, and that had stereotypies higher than the statistical detection threshold, also did not change how variable their locations were (for example, plasma membrane and actin-related structures). All changes in stereotypy in early mitosis were due to a decrease in stereotypy, except for histones, which increased in stereotypy. Together, these observations show that although a concomitant change (or lack of change) in both average location and location variability dominated for most structures during early mitotic reorganization, these two distinct aspects of an individual structure's reorganization were separable for some cellular structures.

We next analysed changes in the relative pairwise locations and their variability in early mitosis through the concordance (Fig. 5d and Extended Data Fig. 10d–f). We found that structures that maintained their locations and their stereotypies also maintained their concordance when paired with each other. Another 64 of the possible 300 pairs of structures changed concordance during early mitosis and these changes were highly linked to changes in stereotypy (Fig. 5d): 61/64 pairs of structures changed in concordance at the same time that at least one of the two structures also changed in stereotypy. For example, the three cytoplasmic organelles, the mitochondria, lysosomes and Golgi, all changed in stereotypy at m2 and all

changed concordance with each other at m2. In 36 of these cases a concordance change occurred at the time of the first stereotypy change of at least one of the two structures (Fig. 5d). For example, the time of the first stereotypy change for mitochondria and microtubules was at m1 because that was when the microtubules changed stereotypy, whereas the mitochondria did not change stereotypy until m2. However, the concordance between this pair of structures already changed at m1, along with the microtubules (and then further at m2, making the concordance stepwise). These results suggest a strong—but not exclusive—relationship between changes in average location, stereotypy, and concordance for many cellular structures during early mitotic reorganization. For 4 out of 64 cases, concordance and stereotypy changed independently for at least one time point (Fig. 5d). Most notable were the histones and microtubules, both of which are central to early mitotic reorganization. For both of these structures, their stepwise ALP was accompanied by a change in stereotypy from interphase to m1 and then a change in concordance from m1 to m2, demonstrating that stereotypy and concordance measurements are separable even for the same pair of structures at two different stages of mitosis.

We performed a meta-analysis to examine all of the possible combinations of the distinct measurements of cell organizational changes used in this analysis framework (Fig. 5e). In this study, all observed examples of changes in any aspect of intracellular organization included a change in the average location of individual cellular structures. Furthermore, in most cases, a change in the relative locations and variability of pairs of structures (concordance) was associated with a change in the variability of at least one of the structures (stereotypy). However, this association between stereotypy and concordance was not absolute, as exemplified by the behaviour of DNA and microtubules in early mitosis.

Discussion

In summary (Extended Data Fig. 11), in this study we introduced the WTC-11 hiPSC Single-Cell Image Dataset v1 and used this resource dataset to develop an analysis framework for integrated intracellular organization. We applied this analysis framework to a large baseline population of cells in interphase, as well as to two subpopulations of cells in the dataset, cells at

the edges of colonies and cells in early mitosis. The results of the meta-analysis investigating the association between distinct aspects of cell organization observed throughout this study suggest a possible hierarchy of dependencies as cells reorganize: (1) the average location of an individual structure changes; (2) the variability in that structure location changes but only when the structure location changes; and (3) the interactions with other structures change, but only when location and/or location variability change. However, our observations also show that this simple proposed hierarchy among these distinct aspects of organization is not absolute—the stereotypy and concordance changed independently in several examples, including for two of the primary structures responsible for early mitotic reorganization, the DNA and the microtubules. It is possible that these potential dependencies, or 'rules' of cell organization, are general and apply to a range of genetic perturbations, differentiation, signalling factors, environmental signals and so on. It is also possible that there is a larger set of cell-type or state-dependent organizational rules.

Together, the raw image data of cells in the dataset, the visualizations and reconstructions of the average locations of cellular structures among the three subsets of cells in this study and the data visualizations constitute a rich resource for further discovery and hypothesis generation. The conceptual aspect of this analysis framework is generalizable and extensible; the establishment of the two conceptual coordinate systems and their application to perform robust statistical analyses on cell shape and intracellular spatial locations and their variability could be useful across different cell types and different types of cell population comparisons. The experimental and algorithmic implementations of this analysis framework are modular, and the choice of which to use is dependent on the specific application. We have demonstrated one specific application to one particular cell type, the hiPS cell, a karyotypically normal cell-culture model system that grows in epithelial-like colonies with a mostly consistent appearance, including an assessment of the required number of cells for these analyses (Supplementary Methods). The specific biological question, cell type or application will dictate the specific inputs required, such as how many cells or cellular structures are needed, what kind of precision is possible or what kinds of segmentation and data-analysis algorithms should be used.

Other systematic image-based approaches have catalogued the location of human proteins in several cell types and used the locations of proteins and structures within cells to identify differences in intracellular spatial patterns among cells in distinct states[6,7,8,9,10,11,12,19,29]. Our work complements these approaches with its focus on analyses of 3D cell organization at the intermediate level of cellular structures (rather than individual proteins), and on the generation of quantitative measurements of distinct aspects of organization, which enables statistical comparisons and provides a more nuanced, systematic definition of cellular organization and reorganization. Together, these studies bring a crucial missing dimension—that is, the spatiotemporal component—to the single-cell revolution[30]. The full image dataset and analysis algorithms introduced here, as well as all the reagents, methods, and tools needed to generate them, are shared in an easily accessible way (https://www.allencell.org/). These data are available to all for further biological analyses and as a benchmark for the development of tools and approaches moving towards a holistic understanding of cell behaviour.

Methods

Cell lines, cell culturing and quality control

Each gene-edited cell line was created using the parental WTC-11 hiPS cell line[31] and contains a fluorescent protein endogenously tagged to a protein representing a distinct cellular structure (Fig. 1a). Cell lines were generated using CRISPR–Cas9-mediated genome editing[14] . The tagging strategy for AAVS1 safe harbour targeting was altered for expression of CAAX-mTagRFP-T[32,33]. Fifteen additional Allen Cell Collection lines were generated using the same methods. The complete list of cell lines and reagents can be found in Supplementary Data 2. The cell lines are described at https://www.allencell.org/cell-catalog.html and are available through Coriell at https://www.coriell.org/1/AllenCellCollection. For all non-profit institutions, detailed MTAs for each cell line are listed on the Coriell website. Please contact Coriell regarding for-profit use of the cell lines as some commercial restrictions may apply. All cell lines were cultured on an automated cell-culture platform developed on a Hamilton Microlab STAR

Liquid Handling System (Hamilton Company). Cells were cultured in a Cytomat 24 (Thermo Fisher Scientific) at 37 °C and 5% CO_2 in mTeSR1 medium with and without phenol red (STEMCELL Technologies), supplemented with 1% penicillin–streptomycin (Thermo Fisher Scientific). Cells were passaged every four days as single cells for up to ten passages post-thaw. For imaging, cells were plated on Matrigel-coated glass-bottom, black-skirt, 96-well plates with 1.5 optical grade cover glass (Cellvis). Cells were regularly assessed for morphology, cell stemness marker expression and outsourced cytogenetic analyses throughout the three years of data acquisition of the WTC-11 hiPSC Single-Cell Image Dataset v1 (ref. [34]). Standard protocols are available at https://www.allencell.org/. Further details are provided in the Supplementary Methods.

Microscopy

Imaging was performed on three identical ZEISS spinning-disk confocal microscopes with 10×/0.45 NA Plan-Apochromat or 100×/1.25 W C-Apochromat Korr UV Vis IR objectives (Zeiss) and ZEN 2.3 software (blue edition; ZEISS) unless otherwise specified. The spinning-disk confocal microscopes were equipped with a 1.2× tube lens adapter for a final magnification of 12× or 120×, respectively, a CSU-X1 spinning-disk scan head (Yokogawa) and two Orca Flash 4.0 cameras (Hamamatsu). Standard laser lines were used at the following laser powers measured with 10× objectives; 405 nm at 0.28 mW, 488 nm at 2.3 mW, 561 nm at 2.4 mW and 640 nm at 2.4 mW unless otherwise specified. An Acousto-Optic Tunable Filter (AOTF) was used to simultaneously modulate the intensity of the four laser lines. The following Band Pass (BP) filter sets (Chroma) were used to collect emission from the specified fluorophore: 450/50 nm for detection of DNA dye, 525/50 nm for detection of mEGFP tag, 600/50 nm for detection of mTagRFP-T tag and 706/95 nm for detection of cell-membrane dye. Images were acquired with an exposure time of 200 ms unless otherwise specified. Cells were imaged in phenol red-free mTeSR1 medium on the stage of microscopes outfitted with a humidified environmental chamber to maintain cells at 37 °C with 5% O_2 during imaging. Transmitted light (bright-field) images were acquired using a white LED light source with broad emission spectrum (pipeline 4.0–4.2) or a red LED light source with peak emission of 740 nm with narrow range and a BP filter 706/95 nm for

bright-field light collection (Pipeline 4.4 only). A Prior NanoScan Z 100 mm piezo z stage (ZEISS) was used for fast acquisition in z (Pipeline 4.4 only). Optical control images were acquired daily at the start of each data acquisition to monitor microscope performance. Laser power was measured monthly and the corresponding percentage adjusted accordingly for each wavelength.

Image acquisition

The image acquisition workflow and experimental set-up evolved over the three years of dataset collection and was versioned into four pipelines. Adjustments included single versus dual camera, filter and light sources, as well as addition of a photoprotective cocktail (Supplementary Methods and Extended Data Fig. 1d). Low magnification (12×), 2D bright-field overview images of cells in wells were collected for cell morphology assessment and for selection of imaging positions for high-magnification (120×), 3D, multichannel imaging. Cells were imaged in three modes to acquire a variation of locations within hiPS cell colonies. Selection of FOV position was performed manually using the stage function in ZEN software or using an automated method, depending on the mode and the cell line. After the selection of FOV position from the well overview acquisition, the DNA of cells was first stained for 20 min with NucBlue Live (Thermo Fisher Scientific). Then the cell membrane was stained with CellMask Deep Red (CMDR, Thermo Fisher Scientific) in the continued presence of NucBlue Live for an additional 10 min, and cells were washed once before imaging for a maximum of 2.5 h. Three-dimensional FOVs at 120× were acquired at the pre-selected positions. Four channels were acquired at each *z*-step (interwoven channels) in the following order: bright field, mEGFP or mTagRFP-T, CMDR and NucBlue Live. Further details are provided in the Supplementary Methods.

3D FOV image quality control

FOV images acquired with two cameras underwent a channel alignment procedure. All 3D FOV images underwent an image quality-control procedure, including three automated FOV quality-control steps. Typical FOV exclusion criteria were related to microscope acquisition system

failures (laser, exposure time, *z*-slice positioning in relation to cell height, empty or out of order channels), analysis steps to identify outliers or any other issues that would cause downstream processing, such as cell, nuclear and cellular structure segmentation, to fail in a systematic batch manner. Total days of acquisition and FOV number per cellular structure are provided in Supplementary Data 1. Further details are provided in the Supplementary Methods.

3D cell and nuclear segmentation

To segment each individual cell and its corresponding DNA from the membrane dye and DNA dye channels of each 3D *z*-stack, we used the deep-learning-based cell and nuclear instance segmentation algorithm developed as part of Allen Cell & Structure Segmenter, an open-source, Python-based 3D segmentation software package[15]. We combined the Segmenter's Iterative Deep Learning workflow and the Training Assay approach to ensure accurate and robust segmentation at scale (18,100 FOVs) for downstream quantitative analysis. We manually validated a subset of the cell and nuclear segmentation results and found that over 98% of individual cells were well-segmented and over 80% of images generated successful cell and nuclear segmentations for all cells in the entire FOV. On the basis of these validation results, we decided that the cell and nuclear instance segmentation algorithm was sufficiently reliable to be applied to all of the FOVs in the dataset. In addition, all cells in the final dataset were manually reviewed for basic quality criteria. Further details are provided in the Supplementary Methods.

3D cellular structure segmentation

We applied a collection of modular segmentation workflows from the Classic Segmentation component of the Segmenter, each optimized for the particular morphological features of the target cellular structures[15]. Representative examples for each of the 25 FP-tagged cellular structures are shown in Extended Data Fig. 2. For each structure, the results of the segmentation workflow were evaluated on sets of images representing the variation observed across imaged cells (for example, different regions of colonies) to ensure consistent segmentation quality across all images for

each structure. We performed an additional validation step to determine whether a given target structure segmentation was sufficient for interpretation in the cellular structure volume analysis (Extended Data Fig. 8). We identified ten structures for which there were obvious caveats to the ability to use their target structure segmentation for biological interpretations of how much of the target structure was present in each cell and thus these ten structures were excluded from the structure volume analysis (Extended Data Fig. 2b–d). Further details are provided in the Supplementary Methods.

Single-cell datasets, feature extraction and quality control

To build the WTC-11 hiPSC Single-Cell Image Dataset v1, we extracted all complete individual cells in each FOV automatically from the cell segmentation results (around 12 complete cells per FOV, on average). All images were rescaled to isotropic voxel size (0.108333 µm in *x*,*y* and *z*). A cropping region of interest (ROI) was created for each cell and applied to each of the original intensity *z*-stacks and cell, nuclear and structure segmentations. Features that were calculated for each cell included FOV-based features (for example, the lowest and highest *z* position of all cells in the FOV), colony-based features (for example, size of the colony), single-cell-based features (for example, cell, nuclear, and cellular structure volume), and single-cell deep-learning-based annotations of cell-cycle stage (for example, interphase or mitotic). The baseline interphase dataset was created by removing all of the 11,190 mitotic cells, as well as approximately 0.5% of outlier cells. We performed an extensive analysis to identify and account for any potential experimental contributions to cell-shape variation (Extended Data Fig. 12). All of the results together confirmed that although cell line identity can contribute to variation in cell height because each cell line was imaged under a particular set of imaging conditions, which varied throughout the imaging pipeline timeline, cell line identity itself does not greatly contribute to the variation in cell height observed in the baseline interphase dataset. Total numbers of cells per cellular structure and per dataset can be found in Extended Data Fig. 1d and Supplementary Data 1. Further details are provided in the Supplementary Methods.

SHE of cell and nuclear shapes

We used SHE coefficients as shape descriptors for cell and nuclear shape[18,35]. We created a publicly available Python package, aics-shparam (see Code availability) to extract SHE coefficients from segmented images of cells and nuclei. Cells and nuclei were first rotated in the *xy* plane such that the longest cell axis falls along the *x* axis. The *z* axis in the lab frame of reference was preserved as it represents the apical–basal axis of these epithelial-like cells. We expanded, up to degree $L_{max} = 16$, resulting in 289 coefficients for each input. Therefore, the shape of each cell in our dataset can be represented by a total of 578 coefficients (Fig. 2a). We could also do the reverse and recreate the 3D mesh representation of a particular set of SHE coefficients with aics-shparam. Further details are provided in the Supplementary Methods.

Building the cell and nuclear shape space

We used PCA to reduce the dimensionality of our joint vectors for all cells (578 SHE coefficients) down to eight principal components. We used the PCA implementation from the Python library scikit-learn[36] with default parameters (Fig. 2b). Because the sign of a given PC is arbitrary, we adjusted the signs where needed to match the naming of the shape modes (for example, larger cells have a more positive PC). We also translated the location of the nuclear mesh back to its correct location relative to the centre of the cell. To prevent cells with extreme shapes from affecting the interpretation of the PCs, we excluded all cells that fell into the range 0th to 1st or 99th to 100th percentiles of each PC from subsequent analysis (remaining $n = 175{,}147$ cells) We *z*-scored all PCs independently by dividing the PC values by the standard deviation (σ) of that PC. The combination of the first eight 'shape modes' (*z*-scored PCs) created the 8D shape space. We used the inverse of the PCA transform generated above to map coordinates from the shape space back into SHE coefficients, which, in turn, were used to reconstruct the corresponding 3D shape. For example, the eight-component vector (0,0,0,0,0,0,0,0) represents the origin of the shape space and its corresponding 3D shape is called the 'mean cell and nuclear shape' (Fig. 2c). In addition to the joint cell and nuclear shape space, we also generated independent cell-only and nucleus-only shape spaces for the baseline interphase dataset (Extended Data Fig. 3e–f), a joint cell and nuclear shape space for cells located at the edges of hiPS cell colonies, and

one each joint cell and nuclear shape space for cells in prophase and in early prometaphase. Finally, we created three joint cell and nuclear shape spaces for the three shape-matched datasets described below. Further details are provided in the Supplementary Methods.

PILRs

The nuclear centroid of each cell was defined as the SHE coefficients representing a one-pixel radius (0.108 µm) 3D spherical mesh. Then, pre-computed SHE coefficients were interpolated to create a series of successive 3D concentric mesh shells from the centroid of the nucleus to the nuclear boundary and from the nuclear boundary to the cell boundary. The *xyz* coordinates of points in the 3D meshes map to corresponding *xyz* locations in the aligned segmented images that were used to generate the SHE coefficients in the first place. Thus, the presence or absence of a segmentation result at each mesh *xyz* coordinate could be organized as a matrix as shown in Fig. 3b. This matrix encodes a PILR of the cell. This process could also be performed using the intensity value at a given *xyz* location in the original FP image (Extended Data Fig. 4). A PILR could then be used to map the cellular structure locations from one cell and nuclear shape into the equivalent locations in any other cell and nuclear shape, thus generating a 'morphed cell' and its reconstructed image. Further details are provided in the Supplementary Methods.

Integrating average morphed cells in the mean cell and nuclear shape

We identified and grouped a set of cells by their absolute proximity in 8D space to the origin of the shape space, map point (0,0,0,0,0,0,0,0). We determined the radius of a sphere centred at this origin such that the number of cells per structure within this sphere was as similar as possible to the average number of cells found in the centre bins of all of the shape modes. A total of 35,633 cells across all 25 structures were found to be within this radius of 2.1σ (see Supplementary Data 1 for numbers of cells per structure). We computed the average of all the PILRs for each structure for all cells within the 8-dimensional sphere. We then morphed these average PILRs into the mean cell and nuclear shape, creating an integrated average morphed

cell. Any cellular structures could be rendered simultaneously to illustrate the spatial relationships of different structures on the basis of their average location in cells of a particular shape.

Pairwise average interaction map of cellular structures

We calculated the 2D pixel-wise Pearson correlation between the averaged PILRs for all pairs of cellular structures within the 8-dimensional sphere, representing a measure of the average location similarity between two structures (Extended Data Fig. 4g). All correlation values used throughout this paper were calculated using the function corrcoef from the Python package NumPy[37]. The average location similarities were organized in a 25 × 25 matrix that represents an average pairwise spatial interaction map of cellular structures (Fig. 3d). This correlation matrix was used as input for a hierarchical clustering algorithm to cluster all 25 cellular structures according to their average location similarities. We used the function cluster.hierarchy.linkage of type 'average' from the Python package scipy[38] to produce the clustering represented by the dendrogram in Fig. 3d. We also computed the average location similarity for every map point along each shape mode. For a given map point, the correlations were computed between the averaged PILRs over all cells that fall into the corresponding map point bin. The heat maps of the resulting matrices for all shape modes and bins between -2σ and 2σ are shown in Fig. 3e and Extended Data Fig. 4h and the data can be found in Supplementary Data 1.

Location stereotypy and location concordance

We calculated the 2D pixel-wise Pearson correlation between the PILRs for all pairs of individual cells within the 8-dimensional sphere centred at the origin of our shape space. This computation results in a 35,633 × 35,633 correlation matrix (Extended Data Fig. 6a). Correlation values from this matrix were averaged within each pair of structures to create an average correlation matrix. Two distinct measurements of structure location and its variation were derived from this average correlation matrix. The diagonal values are the location stereotypy of a given structure and the off-diagonal values are the location concordance between two structures (Extended Data Fig. 6b). We also computed the average correlation matrices for every map

point along each shape mode. For a given map point, the correlations were computed between PILRs over all cells that fall into the corresponding map point bin and then averaged. Heat maps and values of location stereotypy and location concordance for all shape modes and map points can be found in Extended Data Figs. 6c,d and 7c,d and Supplementary Data 1.

Shape-matched datasets

To compare a second, distinct population of cells, such as cells at the edges of colonies or cells in early mitosis, with the baseline interphase cell dataset we created shape-matched datasets. We first mapped cell and nuclear shapes from the second population into the shape space of the baseline dataset by transforming the SHE coefficients from the second population using the same PCs obtained for the baseline dataset. Here we did not exclude cells that fell into the range 0th to 1st or 99th to 100th percentiles of each PC in the baseline dataset because these cells could have shapes more similar to the second population. We then calculated the distance in 8D shape space between every possible pair of cells in both datasets (Extended Data Fig. 9a). Finally, for every cell in the second dataset, we flagged its nearest neighbour within the baseline dataset. The same cell in the baseline dataset could be flagged more than once for multiple different cells within the second dataset. This occurred roughly 12% of the time. The resultant shape-matched dataset is the set of unique flagged cells in the baseline dataset combined with cells in the second dataset. The mean cell shape of this shape-matched dataset is the cell and nuclear shape corresponding to the origin of the corresponding shape-matched shape space. Further details are provided in the Supplementary Methods.

LDA

We performed a PCA dimensionality reduction on all of the PILRs for a given cellular structure in a given shape-matched dataset. This reduced the initial dimensionality of 532,610 pixels in each PILR down to 32 dimensions (or the total number of cells available if fewer than 32). The dimensionally reduced data were then used as input for a LDA to identify the linear combination of reduced dimensions that best separated the two populations of cells within the shape-matched dataset. LDA generates a discriminant

axis along which we could reconstruct corresponding PILRs using the inverse of the PCA transform (Extended Data Fig. 9c and Supplementary Methods). These PILR reconstructions were morphed into the mean cell and nuclear shape for that shape-matched dataset (for example, Supplementary Videos 4 and 5). These reconstructions represent the full range of the ALP for that structure. Each cell was also assigned a location along the discriminant axis (for example, histograms in Extended Data Fig. 9h and Supplementary Videos 4 and 5).

Workflow to flag significant changes in location stereotypy and concordance in early mitosis

To flag whether a difference in location stereotypy or concordance was significant, we first set a threshold cut-off value of Pearson correlation $\rho = 0.03$, below which a stereotypy or concordance value was too low to be used for the subsequent detection of a difference between the baseline dataset and its shape-matched comparison dataset. Next, we set a cut-off threshold for the Pearson correlation value of the difference (ρ_{diff}) in stereotypy or concordance of $\rho_{diff} = 0.02$ (Supplementary Methods). We next applied this workflow to flag all entries in the three early mitotic average correlation difference matrices that showed a significant change between interphase, prophase and early prometaphase (i1–m1, i2–m2 and m1–m2). The first cut-off, $\rho = 0.03$, was applied to the interphase cells when comparing to each early mitotic (i1 for i1–m1; i2 for i2–m2) and to prophase when comparing between the two early mitotic stages (m1 for m1–m2) as in Fig. 5c and Extended Data Fig. 10f. This flagging procedure resulted in three binarized versions of the matrix, in which each flagged entry is marked in black. The combined pattern of flags in these three matrices permits us to identify the TOC for each of the flagged entries (Fig. 5c,d). The four TOC categories included: (1) m1-only: changes that occurred from interphase to m1 but not any further in m2; (2) stepwise: changes that occurred both from interphase to m1 and from m1 to m2; (3) m2-change: changes that occurred from m1 to m2 only; and (4) no change or cases for which changes could not be determined for technical reasons (Fig. 5b and Supplementary Methods). We used all possible combinations of the TOC for the two stereotypies and single concordance for each pair of structures to assess the overall relationship between stereotypy and concordance in early mitosis, which we

consolidated and summarized into three categories (top triangle; Fig. 5d and Supplementary Methods).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets generated during this study, including FOVs, single-cell images and 12× colony overviews, are available at Quilt as packages. Supplementary Data 1 contains (1) a summary of all of the numbers of FOVs, imaging days and cells for all analyses; (2) the correlation values used to generate the heat map data for the average location similarities, stereotypy and concordance, including difference heat maps; and (3) additional data on the comparative analysis of cellular structure volumes in edge and non-edge cells. The full dataset is available at https://open.quiltdata.com/b/allencell/packages/aics/hipsc_single_cell_image_dataset. The dataset containing the non-edge cells shape-matched to edge cells is available at https://open.quiltdata.com/b/allencell/packages/aics/hipsc_single_nonedge_cell_image_dataset. The edge cells dataset is available at https://open.quiltdata.com/b/allencell/packages/aics/hipsc_single_edge_cell_image_dataset. The interphase cells (i1) shape-matched to prophase cells (m1) dataset is available at https://open.quiltdata.com/b/allencell/packages/aics/hipsc_single_i1_cell_image_dataset. The prophase dataset (m1) dataset is available at https://open.quiltdata.com/b/allencell/packages/aics/hipsc_single_m1_cell_image_dataset. The dataset containing the interphase cells (i2) shape-matched to early-prometaphase cells (m2) is available at https://open.quiltdata.com/b/allencell/packages/aics/hipsc_single_i2_cell_image_dataset. The early-prometaphase dataset (m2) dataset is available at https://open.quiltdata.com/b/allencell/packages/aics/hipsc_single_m2_cell_image_dataset. The 12× colony dataset is available at https://open.quiltdata.com/b/allencell/packages/aics/hipsc_12x_overview_image_dataset. The supplementary MYH10 repeat dataset is available at

https://open.quiltdata.com/b/allencell/packages/aics/hipsc_single_cell_image_dataset_supp_myh10. The supplementary training set of 5,664 cells used to train the single-cell classifier is available at https://open.quiltdata.com/b/allencell/packages/aics/mitotic_annotation. The Cell Feature Explorer—215,081 cells (from 18,100 FOVs); 25 structures; 10 features ± apical and radial proximity is available at https://cfe.allencell.org. Source data are provided with this paper.

Code availability

Custom codes were central to the conclusions of the paper. All necessary code to reproduce the results in this paper has been deposited in GitHub. This includes code for downloading our datasets, single-cell feature extraction, cellular parameterization and organelle size scaling. Jupyter notebooks to reproduce the figures shown in the paper are also provided. The released custom code repositories use the following Python packages in parts: NumPy[37] v.1.21.5, Scipy[38] v.1.7.3, scikit-image[39] v.0.19.1, scikit-learn[36] v.1.0.1, Seaborn[40] v.0.11.1, PyTorch[41] v.1.0.0, PyTorchLightning[42] v.0.7.6, VTK[43] v.9.0.1, ITK[44] v.5.2.0, pandas[45] v.1.3.5, matplotlib[46] v.3.5.1, aicsshparam v.0.1.1, aicscytoparam v.0.1.6, pyshtools[47] v.4.9.1, actk v.0.2.2 and aicsimageio[48] v.3.3.2 and v.4.1.0. We also use the softwares: R Statistical Software[49] v.2022.02.2+485, napari[50] v.0.2.8, ChimeraX[51] v.1.3, the Allen Cell & Structure Segmenter[15] (aicssegmentation v.0.1.20, aicsmlsegmentation v.0.0.7, segmenter-model-zoo v.0.0.5) and label free[52] (see below for version). Tutorials and a demo for how to access the data for different purposes are available at https://github.com/AllenCell/quilt-data-access-tutorials. The main codebase used in this paper provides functions for computing features, shape space, shape modes, stereotypy, concordance and morphed cells. The repository also contains the notebooks used to generate the figures shown in the paper. This codebase is available at https://github.com/AllenCell/cvapipe_analysis. The code for shape parameterization via spherical harmonics is available at https://github.com/AllenCell/aics-shparam. The code for cellular parameterization is available at https://github.com/AllenCell/aics-cytoparam. The code for organelle size-scaling analysis is available at https://github.com/AllenCell/stemcellorganellesizescaling. The mitotic

image classifier code[35,40], (for both training and testing) and all trained models is available at https://github.com/AllenCell/image_classifier_3d.The segmentation code used to reproduce the deep learning cell and nuclear segmentations, trained models and demo Jupyter notebook is available at https://github.com/AllenCell/segmenter_model_zoo.The segmentation code used to reproduce structure segmentation from a set of algorithms to choose from, each with restricted numbers of parameters to tune, is available at https://github.com/AllenCell/aics-segmentation.The code used to generate the contact sheet quality-control single-cell visualizations of all segmented cells is available at https://github.com/AllenCellModeling/actk. The code to create the 12× colony dataset is available at https://github.com/AllenCell/colony-processing. The customized label-free code used as part of the cell and nuclear segmentation model is available at https://github.com/AllenCellModeling/pytorch_fnet/tree/50c433c2e72d2d42886b48c5faf5449725d195a5. Software will be shared under the Allen Institute Software License and Contribution Agreement, subject to any applicable third-party licensing restrictions. Datasets will be shared under the Allen Institute Terms of Use: https://alleninstitute.org/legal/terms-use/.

References

1. Kirschner, M., Gerhart, J. & Mitchison, T. Molecular “vitalism”. *Cell* **100**, 79–88 (2000).

2. Woese, C. R. A new biology for a new century. *Microbiol. Mol. Biol. Rev.* **68**, 173–186 (2004).

3. Karsenti, E. Self-organization in cell biology: a brief history. *Nat. Rev. Mol. Cell Biol.* **9**, 255–262 (2008).

4. Rafelski, S. M. & Marshall, W. F. Building the cell: design principles of cellular architecture. *Nat. Rev. Mol. Cell Biol.* **9**, 593–602 (2008).

5. Roggiani, M. & Goulian, M. Oxygen-dependent cell-to-cell variability in the output of the *Escherichia coli* Tor phosphorelay. *J. Bacteriol.* **197**, 1976–1987 (2015).

6. Caicedo, J. C. et al. Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).

7. Thul, P. J. et al. A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).

8. Cai, Y. et al. Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature* **561**, 411–415 (2018).

9. Gut, G., Herrmann, M. D. & Pelkmans, L. Multiplexed protein maps link subcellular organization to cellular states. *Science* **361**, eaar7042 (2018).

10. Gerbin, K. A. et al. Cell states beyond transcriptomics: integrating structural organization and gene expression in hiPSC-derived cardiomyocytes. *Cell Syst.* **12**, 670–687 (2021).

11. Qin, Y. et al. A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 (2021).

12. Cho, N. H. et al. OpenCell: endogenous tagging for the cartography of human cellular organization. *Science* **375**, eabi6983 (2022).

13. Drubin, D. G. & Hyman, A. A. Stem cells: the new “model organism”. *Mol. Biol. Cell* **28**, 1409–1411 (2017).

14. Roberts, B. et al. Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Mol. Biol. Cell* **28**, 2854–2874 (2017).

15. Chen, J. et al. The Allen Cell Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images. Preprint at *bioRxiv* https://doi.org/10.1101/491035 (2018).

16. Pincus, Z. & Theriot, J. A. Comparison of quantitative methods for cell-shape analysis. *J. Microsc.* **227**, 140–156 (2007).

17. Marshall, W. F., Dernburg, A. F., Harmon, B., Agard, D. A. & Sedat, J. W. Specific interactions of chromatin with the nuclear envelope: positional determination within the nucleus in *Drosophila melanogaster*. *Mol. Biol. Cell* **7**, 825–842 (1996).

18. Ruan, X. & Murphy, R. F. Evaluation of methods for generative modeling of cell and nuclear shape. *Bioinformatics* **35**, 2475–2485 (2019).

19. Schauer, K. et al. Probabilistic density maps to study global endomembrane organization. *Nat. Methods* **7**, 560–566 (2010).

20. Wang, T. & Hong, W. Interorganellar regulation of lysosome positioning by the golgi apparatus through Rab34 interaction with Rab-interacting lysosomal protein. *Mol. Biol. Cell* **13**, 4317–4332 (2002).

21. Hao, F. et al. Rheb localized on the Golgi membrane activates lysosome-localized mTORC1 at the Golgi–lysosome contact site. *J. Cell Sci.* **131**, jcs208017 (2018).

22. Rowland, A. A. & Voeltz, G. K. Endoplasmic reticulum–mitochondria contacts: function of the junction. *Nat. Rev. Mol. Cell Biol.* **13**, 607–625 (2012).

23. Doghman-Bouguerra, M. & Lalli, E. ER–mitochondria interactions: both strength and weakness within cancer cells. *Biochim. Biophys. Acta Mol. Cell Res.* **1866**, 650–662 (2019).

24. Marshall, W. F. Scaling of subcellular structures. *Annu. Rev. Cell Dev. Biol.* **36**, 219–236 (2020).

25. Valencia, P., Dias, A. P. & Reed, R. Splicing promotes rapid and efficient mRNA export in mammalian cells. *Proc. Natl Acad. Sci. USA* **105**, 3386–3391 (2008).

26. Wurm, C. A. et al. Nanoscale distribution of mitochondrial import receptor Tom20 is adjusted to cellular conditions and exhibits an inner-cellular gradient. *Proc. Natl Acad. Sci. USA* **108**, 13546–13551 (2011).

27. Kim, Y. et al. Cell position within human pluripotent stem cell colonies determines apical specialization via an actin cytoskeleton-based mechanism. *Stem Cell Rep.* **17**, 68–81 (2022).

28. Champion, L., Linder, M. I. & Kutay, U. Cellular reorganization during mitotic entry. *Trends Cell Biol.* **27**, 26–41 (2017).

29. Donovan-Maiye, R. M. et al. A deep generative model of 3D single-cell organization. *PLoS Comput. Biol.* **18**, e1009155 (2022).

30. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 4307 (2020).

31. Kreitzer, F. R. et al. A robust method to derive functional neural crest cells from human pluripotent stem cells. *Am. J. Stem Cells* **2**, 119–131 (2013).

32. Hockemeyer, D. et al. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat. Biotechnol.* **27**, 851–857 (2009).

33. Oceguera-Yanez, F. et al. Engineering the AAVS1 locus for consistent and scalable transgene expression in human iPSCs and their differentiated derivatives. *Methods* **101**, 43–55 (2016).

34. Coston, M. E. et al. Automated hiPSC culture and sample preparation for 3D live cell microscopy. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.18.423371 (2020).

35. Shen, L., Farid, H. & McPeek, M. Modeling three-dimensional morphological structures using spherical harmonics. *Evolution* **63**, 1003–1016 (2009).

36. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

37. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).

38. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Meth*. **17**, 261–272 (2020).

39. Walt, S. V. D. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

40. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

41. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS)* (eds Wallach, H. et al.) 8026–8037 (NeurIPS, 2019).

42. Falcon, W. et al. PyTorchLightning/pytorch-lightning: 0.7.6 release. https://doi.org/10.5281/ZENODO.3828935 (2020).

43. Schroeder, W., Martin, K. & Lorensen, B. *The Visualization Toolkit: An Object-Oriented Approach To 3D Graphics* (Kitware, 2018).

44. McCormick, M. M., Liu, X., Ibanez, L., Jomier, J. & Marion, C. ITK: enabling reproducible research and open science. *Front. Neuroinform.* **8**, 13 (2014).

45. McKinney, W. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* **445**, 56–61 (SCIPY, 2010).

46. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

47. Wieczorek, M. A. & Meschede, M. SHTools: tools for working with spherical harmonics. *Geochem. Geophys. Geosyst.* **19**, 2574–2592 (2018).

48. Maxfield Brown, E. et al. AICSImageIO: image reading, metadata conversion, and image writing for microscopy images in pure Python. https://pypi.org/project/aicsimageio/ (2021).

49. R Core Team. *R: A Language and Environment for Statistical Computing: Reference Index* (R Foundation for Statistical Computing, 2010).

50. Sofroniew, N. et al. napari/napari: 0.2.8. https://doi.org/10.5281/zenodo.3592005 (2019).

51. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2020).

52. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917–920 (2018).

53. McHugh, M. L. The chi-square test of independence. *Biochem. Med.* **23**, 143–149 (2013).

Acknowledgements

We thank J. Brugge, G. Danuser, M. Elowitz, T. Goddard, Q. Justman, J. Lippincott-Schwartz, W. Marshall, S. Palacek, Z. Pincus and T. Pollard for scientific discussions. The WTC line that we used to create our gene-edited cell lines was provided by the Bruce R. Conklin Laboratory at the Gladstone Institute and UCSF. J.A.T. was supported by the Howard Hughes Medical Institute (HHMI) and the Washington Research Foundation. A.L.S. was supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship. R.L.D.L. was supported by the Stanford Medical Scientist Training Program, National Institutes of Health T32-GM007365. This article is subject to HHMI's Open Access to Publications policy. HHMI laboratory heads have previously granted a nonexclusive CC BY 4.0 licence to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licences, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 licence immediately upon publication. We acknowledge the Allen Institute for Cell Science founder, Paul G. Allen, for his vision, encouragement and support.

Author information

Author notes

1. These authors contributed equally: Jianxu Chen, Theo A. Knijnenburg, Ritvik Vasan, Calysta Yan

2. Deceased: Magdalena Schwarzl

Authors and Affiliations

1. Allen Institute for Cell Science, Seattle, WA, USA

   Matheus P. Viana, Jianxu Chen, Theo A. Knijnenburg, Ritvik Vasan, Calysta Yan, Joy E. Arakaki, Matte Bailey, Ben Berry, Antoine Borensztejn, Eva M. Brown, Sara Carlson, Julie A. Cass, Basudev Chaudhuri, Kimberly R. Cordes Metzler, Mackenzie E. Coston, Zach J. Crabtree, Steve Davidson, Colette M. DeLizo, Shailja Dhaka, Stephanie Q. Dinh, Thao P. Do, Justin Domingus, Rory M. Donovan-Maiye, Alexandra J. Ferrante, Tyler J. Foster, Christopher L. Frick, Griffin Fujioka, Margaret A. Fuqua, Jamie L. Gehring, Kaytlyn A. Gerbin, Tanya Grancharova, Benjamin W. Gregor, Lisa J. Harrylock, Amanda Haupt, Melissa C. Hendershott, Caroline Hookway, Alan R. Horwitz, H. Christopher Hughes, Eric J. Isaac, Gregory R. Johnson, Brian Kim, Andrew N. Leonard, Winnie W. Leung, Jordan J. Lucas, Susan A. Ludmann, Blair M. Lyons, Haseeb Malik, Ryan McGregor, Gabe E. Medrash, Sean L. Meharry, Kevin Mitcham, Irina A. Mueller, Timothy L. Murphy-Stevens, Aditya Nath, Angelique M. Nelson, Sandra A. Oluoch, Luana Paleologu, T. Alexander Popiel, Megan M. Riel-Mehan, Brock Roberts, Lisa M. Schaefbauer, Magdalena Schwarzl, Jamie Sherman, Sylvain Slaton, M. Filip Sluzewski, Jacqueline E. Smith, Youngmee Sul, Madison J. Swain-Bowden, W. Joyce Tang, Derek J. Thirstrup, Daniel M. Toloudis, Andrew P. Tucker, Veronica Valencia, Winfried Wiegraebe, Thushara Wijeratna, Ruian Yang, Rebecca J. Zaunbrecher, Graham T. Johnson, Ruwanthi N. Gunawardane, Nathalie Gaudreault & Susanne M. Rafelski

2. Department of Biology and Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

Ramon Lorenzo D. Labitigan & Julie A. Theriot

3. Department of Biochemistry, Stanford University, Stanford, CA, USA

Ramon Lorenzo D. Labitigan

4. Department of Computer Science, Stanford University, Stanford, CA, USA

Adrian L. Sanborn

5. Department of Structural Biology, Stanford University, Stanford, CA, USA

Adrian L. Sanborn

Contributions

E.M.B., J.A.C., J.C., T.P.D., M.A.F., N.G., K.A.G., B.W.G., R.N.G., A.H., M.C.H., C.H., A.R.H., G.T.J., R.L.D.L., J.J.L., A.N., A.M.N., L.P., S.M.R., M.M.R.-M., B.R., A.L.S., L.M.S., M.S., J.S., J.E.S., J.A.T., D.J.T., D.M.T., A.P.T., V.V., M.P.V., W.W. and C.Y. contributed to the development and/or design of the methods used in this paper. M.B., E.M.B., J.A.C., B.C., J.C., Z.J.C., S. Davidson, S. Dhaka, T.P.D., R.M.D.-M., T.J.F., G.F., T.G., L.J.H., H.C.H., E.J.I., G.T.J., G.R.J., B.K., J.J.L., B.M.L., G.E.M., S.L.M., K.M., A.N., L.P., T.A.P., M.M.R.-M., L.M.S., M.S., J.S., S.S., M.F.S., M.J.S.-B., D.J.T., D.M.T., R.V., M.P.V., W.W., T.W., C.Y. and R.Y. contributed to the software that was developed to create, assess and analyse the dataset. J.E.A., A.B., J.A.C., J.C., M.E.C., S.Q.D., M.A.F., N.G., K.A.G., T.G., B.W.G., A.H., C.H., J.J.L., H.M., I.A.M., A.N., A.M.N., L.P., S.M.R., M.M.R., B.R., L.M.S., J.E.S., D.J.T., D.M.T., A.P.T., V.V., M.P.V., C.Y. and R.J.Z. contributed to validating the cell lines used, the images taken and the data herein to ensure reproducibility. J.E.A., E.M.B., J.A.C., S.Q.D., R.M.D.-M., A.J.F., M.A.F., K.A.G., T.G., R.N.G., A.H., G.R.J., T.A.K., J.J.L., H.M., S.A.O., L.P., S.M.R., M.M.R.-M., B.R., L.M.S., J.S., Y.S., D.M.T., A.P.T.,

V.V., R.V., M.P.V., C.Y. and R.J.Z. contributed to the formal analysis that is shown in the main figures and supplementary materials. J.E.A., A.B., S.C., M.E.C., C.M.D., S.Q.D., M.A.F., N.G., J.L.G., K.A.G., T.G., B.W.G., A.H., M.C.H., C.H., W.W.L., S.A.L., H.M., I.A.M., A.N., A.M.N., L.P., S.M.R., B.R., J.E.S., W.J.T., J.A.T., D.J.T., A.P.T., V.V., M.P.V., C.Y., R.Y. and R.J.Z. contributed to collecting the data, performing the experiments or analysing the imaging data. B.B., E.M.B., J.A.C., J.D., M.A.F., B.W.G., A.H., A.N.L., R.M., T.L.M.-S., M.M.R., B.R., J.S., J.E.S., D.M.T., W.W. contributed to providing the reagents, materials or computing resources needed to accomplish this study. J.E.A., M.B., B.B., A.B., E.M.B., S.C., J.C., M.E.C., C.M.D., S.Q.D., R.M.D., T.J.F., M.A.F., N.G., J.L.G., K.A.G., T.G., B.W.G., A.H., M.C.H., C.H., T.A.K., W.W.L., J.J.L., S.A.L., K.M., I.A.M., A.N., M.M.R.-M., B.R., J.S., S.S., J.E.S., M.J.S.-B., W.J.T., D.J.T., D.M.T., A.P.T. and C.Y. contributed to management activities to annotate, scrub data and maintain research data. J.C., M.E.C., T.P.D., N.G., C.H., A.R.H., G.T.J., T.A.K., A.N., S.M.R., J.A.T., R.V., M.P.V. and C.Y. helped to prepare, create and present the original draft of this work. A.B., J.C., K.R.C.M, T.P.D., C.L.F., N.G., C.H., A.R.H., G.T.J., T.A.K., I.A.M., S.M.R., J.A.T., R.V., M.P.V. and C.Y. helped with the reviewing and editing of this work. J.A.C., K.R.C.M, T.P.D., N.G., G.T.J., T.A.K., M.M.R.-M., D.M.T., R.V. and M.P.V. helped to prepare the visualization of the work described, specifically through visual tools, resources and data presentation. J.E.A., K.R.C.M., M.A.F., N.G., K.A.G., R.N.G., A.H., A.R.H., G.T.J., T.A.K., I.A.M., A.M.N., S.M.R., B.R., D.M.T. and W.W. provided oversight and leadership, including the planning and execution of the cross-team project management. K.R.C.M., R.N.G., A.H., G.T.J., I.A.M., S.M.R., S.S. and D.M.T. provided management, coordination of the project and administration of resources.

Corresponding author

Correspondence to Susanne M. Rafelski.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks George Daley, Guillaume Jacquemet and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Creation of the WTC-11 hiPSC Single-Cell Image Dataset v1 that contains over 200,000 live, high-resolution, 3D cells spanning 25 cellular structures.

The dataset was generated by a microscopy pipeline composed of three main parts; Data Collection, Image Processing and Single-Cell Feature Extraction. **a**. Data Collection: the sample preparation starts with a vial of frozen gene-edited hiPS cells from a line from the Allen Cell Collection,

expressing an endogenous, fluorescently tagged protein representing a particular cellular structure. The cells are cultured in 6-well plates on an automated cell-culture platform. At each passage cells are seeded into optical grade, glass-bottom 96-well plates to create imaging samples. Bright-field overview images of each well are inspected and only wells meeting pre-determined quality controls are passaged from the 6-well plates and imaged from the 96-well plates. The image acquisition of live cells starts with a 12X overview image of each well on a spinning-disk confocal microscope to keep track of the position of each image within each colony. Imaging sessions are conducted using three modes to capture variations in colony area, locations within the colony, and enrich for images with mitotic cells as needed. In mode A, the 12X overview images of colonies are segmented by an automated script to generate sets of coordinates for positions within imageable colonies, located approximately halfway between the colony edge and colony centre. Imageable colonies are those that meet size, morphology, and position-within-a-well criteria. In mode B, the microscope operator adjusts the location of the field of view (FOV) to enrich for mitotic cells via appropriate cell and DNA morphology visible with live bright-field viewing and confirmed by DNA staining (yellow arrows). In mode C, three regions of colonies are imaged, the edge, ridge (just inward from the edge), and centre. The combination of these three imaging modes permitted sampling across all regions of the hiPS cell colonies (Extended Data Fig. 12). Cells were labelled with fluorescent DNA and membrane dyes and then imaged at each pre-selected colony position. Z-stacks were acquired at 120X in four channels, representing the bright-field, cell-membrane dye (magenta), DNA dye (cyan) and the fluorescently tagged cellular structure (grayscale), also shown in (b). Mode A and C panels show Golgi (via sialyltransferase) and microtubules (via alpha-tubulin), respectively. **b**. Image Processing: The WTC-11 hiPSC Single-Cell Image Dataset v1 consists of a total of 18,100 FOVs curated specifically for successful cell and cellular structure segmentations, which are available for download. An example z-stack is shown. On the left is the maximum intensity projection of all 65 slices with all fluorescent channels combined, in the colours indicated in the panels on the right. “Cutting” the z-stack in half exposes the view of a single slice (slice 32) in the middle of the stack, shown for each individual channel, including the bright-field channel. We applied 3D segmentation algorithms to each of the fluorescent

channels to identify boundaries in 3D of the cells via the membrane dye (magenta), the nuclei and mitotic DNA via the DNA dye (cyan), and each of the 25 cellular structures via their fluorescent protein tag (grayscale; Golgi shown here). Resulting 3D segmentations for cell membrane, DNA, and structure channels are also shown as a side view, the xz-cross-section along the yellow dotted line. All segmentation algorithms were developed and performed using the Allen Cell & Structure Segmenter. **c**. Single Cell Feature Extraction: A total of 215,081 single cells were segmented from the FOVs. Every individual cell was labelled with a unique ID and metadata related to the sample, experiment, and microscopy was collected and associated with each individual cell for future data provenance. Appropriate features were extracted for each cell from the cell, the nucleus or mitotic DNA, and the cellular structure segmentations, including measurements such as the height and volume. These cells, including the images and the segmentations as well as the metadata and features are all available for download. Scale bars are 10 μm unless otherwise noted. **d**. Number of cells for each cellular structure in the WTC-11 hiPSC Single-Cell Image Dataset v1, sorted by their acquisition order. This table includes all of the various different subsets of the data used throughout the study, including the baseline interphase dataset (excluding outliers, see Methods), mitotic cells, cells within the 8-dimensional sphere (Fig. 3), cells at the edges of colonies (Fig. 4) and cells in early stages of mitosis (m1 and m2, Fig. 5).

Extended Data Fig. 2 Overview of cell, nuclear and cellular structure segmentations and caveats.

a. Panels show a representative single z-slice of the FP-tagged protein (left) and the target segmentation (right), demonstrating the degree of accuracy of the structure segmentations used for analysis. Several of these segmentations have specific types of caveats (**b**–**d** and Methods) that may affect interpretation of downstream analyses. **b**. The limits of the cell boundary segmentation algorithm include potential errors for the very top slices of each cell. Desmosomes, which localize to the cell periphery at the top of the cell, demonstrate this caveat well. Four sequential z-slices (z = 42-48) moving upwards towards the top of the cell-membrane dye signal are shown. In z = 42, both the cell-membrane dye and the cell segmentation clearly identify the true cell boundary (yellow arrows) and in z = 48, the in-

focus desmosomes also line up well along the true cell boundary. However, in z = 46, the cell-membrane dye indicates two possible cell boundaries due to the slanted nature of the top of this cell and the out of focus light spreading from slices above and below. The in-focus desmosomes identify the inner possible boundary as the true cell boundary (yellow arrows). However, the segmented cell boundary is incorrect (cyan arrows). In z = 44 the desmosomes are not yet in focus, thus the true cell boundary is likely somewhere between that determined in z = 42 and z = 46. This error is negligible for overall cell segmentation, but critical for the assignment of desmosome locations in the cell. In this example shown, desmosomes are not located directly at the segmented cell periphery but still close by, such that a measurement of the total volume of desmosomes in this cell is still appropriate. However, it is equally likely that desmosomes, or any other structure localizing to the upper cell periphery could be mis-assigned to a neighbouring cell. Thus these structures were not considered validated for cellular structure volume analyses (Methods). **c**. Structures localizing or partially localizing to a thin 3D surface, such as the cell or nuclear periphery, may suffer from non-uniform accuracy between the middle and the top/bottom of that structure due to the anisotropic resolution of the images. Seven sequential z-slices (left) and target segmentations (right) of nuclear pores on the nuclear surface demonstrate this caveat well. The density of segmented nuclear pores is greatest at z = 36 and declines as the imaging plane moves upward through the nucleus. Consistently accurate detection for nuclear pores at both the centre and the top of the nucleus was not possible due to this effect and would require further algorithm development. The segmentation accuracy was sufficient to identify the general location of nuclear pores in cells for the location-based analyses but not sufficient to be validated for use in the cellular structure volume analysis. This caveat was also observed for other structures localizing to the nuclear and cell periphery (Methods). **d**. The segmentation target for cohesins (via SMC-1A) is to detect the most contrasted locations of cohesins in nuclei. This segmentation works well for nuclei in most of interphase (see example in (**a**)). However, SMC-1A moves from the cytoplasm back into the nucleus after mitosis. The amount of tagged SMC-1A protein in the nucleus and thus its segmentation depends on how far into interphase a cell is. Three examples of tagged SMC-1A are shown (left panels) along with the target segmentations (right panels). For a cell in early

interphase (far left) SMC-1A is both in the cytoplasm and nucleus, but at low levels such that the target segmentation is quite sparse. For a cell well into interphase (far right), the target segmentation is as in (**a**). In the centre is a nucleus with moderate levels of SMC-1A and thus fewer cohesin locations segmented. **e**. Demonstration of the cell membrane Training Assay concept. Top row: tagged cell-membrane channel (via CAAX; left) and cell-membrane dye channel (right) images as single slices near the centre of the z-stack. Second row: corresponding side views of the same z-stacks. Bottom row: CAAX-based segmentation (left) and filled version for the cell-membrane dye-based segmentation performed on the dye image after training via the cell membrane Training Assay (right). The cell membrane at the top of cells is often very dim in the dye images (yellow arrows) due to both the very thin nature of the top membrane and photobleaching during z-stack acquisition. However, the top of these same cells is much more visible in the tagged plasma membrane cell line (cyan arrows), permitting successful CAAX-based segmentations. We leveraged the information contained in the CAAX images by using the CAAX-based segmentation as the training target for a deep learning cell-membrane dye-based segmentation model. **f**. Demonstration of the DNA dye Training Assay concept. Top row: tagged nuclear envelope channel (via lamin B1; left) and DNA dye channel (right) images as a single slice near the centre of the z-stack. Second row: corresponding side views of the same z-stacks. Bottom row: lamin B1-based segmentation (left) and filled version for the DNA dye-based segmentation performed on the dye image after training via the DNA dye Training Assay (right). The top boundaries of nuclei are often very blurry in the DNA dye images (yellow arrows) due to the “filled” nature of how the DNA dye demarcates the nuclear boundary combined with the diffraction of light and lower axial resolution. However, the top boundaries of nuclei in these same cells are clearly identifiable in the tagged nuclear envelope cell line (cyan arrows), permitting accurate nuclear segmentations via lamin B1. We leveraged the image information in the lamin B1 images by using the filled lamin B1-based segmentation as the training target for a deep learning DNA dye-based segmentation model. Total numbers of acquisition days, FOVs, and cells per cellular structure are in Supplementary Data 1 and Extended Data Fig. 1d.Scale bars are 3 μm for **a–d** and 5 μm for **e–f**.

Extended Data Fig. 3 A PCA-based cell and nuclear shape space reveals interpretable modes of shape variation in hiPS cells (supporting figure).

a. Mean distance between points in the original 3D meshes of cell (top) and nucleus (bottom) to their corresponding closest points in the reconstructed meshes and vice versa as the number of coefficients in the SHE increases. Each grey line is one cell (left; n = 300 randomly selected samples) or nucleus (right; n = 300 randomly selected samples). Black lines represent the mean. The dashed vertical lines indicate the number of coefficients for SHE degree $L_{max} = 16$. **b**. Two examples of how nine map points for each of the eight shape modes are used as the input for an inverse PCA transform to obtain the corresponding SHE coefficients and their corresponding 3D reconstructions at these map points. Three 2D views of the 3D shape are shown as in Fig. 2c. The top view corresponds to an intersection between the 3D mesh of the cell and nucleus reconstructions and the xy plane. Side views 1 and 2 correspond to an intersection between the 3D meshes and the *xz*- or *yz*-planes, respectively. **c**. Pairwise correlations for cell volume, cell height, and shape modes. Each point represents a single cell (n = 202,847). Points are colour-coded based on an empirical density estimate. The grey line represents the best linear fit. The green curve represents the non-overlapping window average (y-axis) within 100 equally spaced bins (x axis). Only results for bins with more than 50 points are reported. Pearson correlation values are indicated in the upper triangle part of the figure (black for non-zero values). **d**. Bar graph plots of the total variance explained by each PC for the shape spaces obtained when only nuclear (**e**) and cell (**f**) SHE coefficients are used as input for the PCA dimensionality reduction described in Fig. 2. **e–f**. Most relevant 2D view of 3D shapes reconstructed at each of the nine map points of each of the eight shape modes (given human-interpretable names). The centre bin in all modes is the identical mean cell shape. At the far right is an overlay of 2D views of the nucleus (**e**) or cell (**f**) for the two most extreme map points (at −2σ, lighter shade and +2σ, darker shade) of each shape mode.

Source data

[Extended Data Fig. 4 Creating and comparing integrated average cells throughout the shape space via SHE coefficient-based parameterization and 3D morphing.]

a. "*3-channel original z-stack"* (bottom left image), shows a 3D visualization of the original FP intensities of tagged mitochondria (grayscale) in a single cell and nucleus, visualized via cell-membrane dye (magenta) and DNA dye (cyan). Moving rightward along the bottom row are the steps to create the PILR of the mitochondria via the FP signal in this cell. "*3D reconstruction"* (second image) shows the SHE-based 3D reconstruction meshes of the segmentations of this cell and nucleus. Next, "*cellular mapping"* shows the result of interpolating the SHE coefficients to create a series of successive 3D concentric mesh shells (different colours) from the centroid of the nucleus (black dot) to the nuclear (inner) and then to the cell (outer) boundary to create the nuclear and cytoplasmic mapping, respectively. The intensity values in the FP channel are recorded at each mesh vertex location, resulting in the "*PILR*" that is shown in matrix format in the fourth image. "*Voxelization*" shows the result when this PILR is converted back into a 3D image, voxel by voxel, into the same reconstructed cell and nuclear shape. Because this internal mapping is discrete, the resultant reconstructed image will have gaps. At the top, "*original FP image* (left) is the original image and "*nearest neighbour interpolation"* (right) is the voxelized PILR, now with the gaps filled using nearest neighbour interpolation. Voxel-wise Pearson correlation in 3D is used to compare these original and reconstructed FP images. **b**. Example PILRs (in matrix format as in **a**) for one cell for each of five cellular structures. Top view and side view 1 are shown on the far left. Top and bottom PILR matrices for each structure are based on the original FP image (grayscale on black background) or the structure segmentations (binary on white background), respectively. **c**. The FP-image-based PILR takes all intensities in the image into account, including any FP-tagged protein not localized to the target structure that the protein represents. For example, FP-tagged paxillin localized to matrix adhesions at the bottom of the cell but also throughout the cytoplasm. Two images of multiple cells (cell membrane indicated by magenta lines) in an FOV with labelled matrix adhesions (via paxillin) at two z positions in the z-stack. Top left triangles in each image show the original FP image. Matrix adhesions are visible

near the bottom of the cells (left) but considerable FP-tagged paxillin signal is visible both at the bottom and centre (right) of cells. However, the segmentation target defined for this cell line included only the high intensity regions representing the matrix adhesions near the bottom of the cells. Bottom right triangles in each image show the result of the matrix adhesion specific segmentation. Total numbers of acquisition days, FOVs, and cells for FP-tagged paxillin are in Supplementary Data 1 and Extended Data Fig. 1d. **d**. Using the structure segmentation-based PILR permits the creation of average morphed cells containing the locations of the cellular structures that each tagged protein represents. Average morphed cells representing matrix adhesions (top row) and mitochondria (bottom row) generated using either the original FP images (left column) or the target structure segmentations (right column) of cells within the 8-dimensional sphere morphed into the mean cell shape. The analyses in this paper focus on the structure segmentation-based PILRs; but conceptually the same approach could also be applied to the raw intensity images. **e**. Bar graphs of voxel-wise Pearson correlation between original intensity images of FP-tagged proteins (left) or of structure segmentations (right) and the images reconstructed from the PILR. Error bars represent ± one standard deviation around the mean (n = 32 cells per structure). Cells were selected from centre bin of Shape Mode 1. The correlation for cohesins (via segmentations) is indicated with a striped fill pattern. This structure has a significantly changing target structure segmentation depending on how much tagged cohesin has re-entered the nucleus after mitosis, causing the much lower correlation value (Extended Data Fig. 2d). **f**. Example cell from the top of (**a**) to show the original and PILR-based reconstructed image but here based on the structure segmentations. Numbered insets are zoomed in regions. Cell and nuclear boundaries in **a–f** are shown in magenta and cyan lines, respectively. **g**. Overview of the process to calculate the average location similarity between all pairwise-combinations of the 25 cellular structures within the 8-dimensional shape space sphere. The 2D pixel-wise Pearson correlation was calculated between pairs of averaged PILRs for each structure. This created a correlation matrix including each of the 25 cellular structures with elements of this matrix representing the average location similarity between two cellular structures. **h**. Heat maps for the −2σ and 2σ shape space map points for each of the eight shape modes as in Fig. 3e, but here heat map values correspond to the difference in average structure

similarity between the mean cell shape and either, the −2σ and 2σ shape space map points (bottom and top triangles, respectively), for each of the eight shape modes (numbers of cells in Supplementary Data 1). Due to technical considerations related to the PILR construction (Methods) or due to especially low number of cells in some bins (Supplementary Data 1), some structures displayed changes in the magnitude of the average location similarity with other structures in the shape mode bins furthest from the mean (−2σ and 2σ, mainly for Shape Mode 1) and so these decreases may not be biologically meaningful. Additional difference heat maps for intermediate shape mode bins are available in Supplementary Data 1.

Source data

Extended Data Fig. 5 Overview panel for creating aggregated morphed cells for all 25 cellular structures.

a. Each row represents one of the 25 cellular structures (indicated by the colour bar on the far left). From left to right, on the left side of the large arrow, the first three sets of three images each show top view and side view 1 of three examples of individual cells with shapes similar to the mean cell shape (origin of the 8-dimensional sphere). For each set of three images, the left is the maximum intensity projection (MIP) of the original FP image (grayscale on black background), the centre is the average intensity projection of the structure segmentation image (AIP; binary on white background), and the right is the AIP for the structure segmentation-based PILR for that cell morphed into the mean cell shape. For nuclear envelope and nuclear pores, the centre slice, through the centre of the nucleus, of the original FP image is shown instead of the MIP. For these two structures and for histones, the cyan DNA outline has been left out to see the location of these structures at the nuclear periphery. **b**. On the right side of the large arrow are three different types of aggregations of the indicated number of individual morphed cells based on the structure segmentation PILRs. On the left is the average morphed cell, the centre is the standard deviation (std.) morphed cell, and the right is the "*structure-localized coefficient of variation*" (SLCV) morphed cell, representing a quantitative measure of how variable the location of a structure is at any given voxel (Supplementary Methods). Contrast settings for FP and AIP images were

adjusted per cellular structure to best represent its location. Heat maps for average morphed cells indicate relative likelihood of a structure being at a given location in the cell. Heat map ranges for standard deviation morphed cell and SLCV morphed cell are as described (Supplementary Methods). Scale bars are 5 µm.

Extended Data Fig. 6 Location stereotypy and concordance are robust to systematic variation in cell and nuclear shape.

a. Heat map of the 2D pixel-wise Pearson correlation matrix for all pairs of cellular structure PILRs among all cells in the 8-dimensional sphere. Each entry in this matrix represents the correlation between the PILR of two cells. Coloured triangles to the left of, and the thicker black lines within, the matrix indicate the regions (blocks) of the matrix corresponding to cells with the indicated tagged structure. The dimensions of each block correspond to the number of cells. **b**. Average correlation matrix. Left: the location stereotypy for a cellular structure is the average of all the values in the blocks along the diagonal of the correlation matrix in (a). The numbers on the right indicate structures ranked by their stereotypy from greatest to least. Right: the location concordance for any two pairs of structures is the average of all the values in the corresponding structure pair block in the correlation matrix in (a). The diagonal of the concordance heat map corresponds to the stereotypy. Arrows indicate examples of the relationships between the heat maps in (a) and (b). **c**. Stereotypy heat maps for each of the eight shape modes (SM). Each row represents a different cellular structure and each column represent the nine binned map points along each shape mode (Fig. 2b). **d**. Concordance heat maps for the −2σ and 2σ shape space map points for each of the eight shape modes. The lower and upper triangles represent the −2σ and 2σ map points, respectively. Numbers of cells and heat map data in Supplementary Data 1. Colour bars on the left of heat maps indicate the cellular structure.

Source data

Extended Data Fig. 7 Comparing location stereotypy and concordance throughout the cell and nuclear shape space.

a. Box plots of the diagonal values for each of the 25 cellular structures in the 3D voxel-wise Pearson correlation matrix heat map for all cells in the 8-dimensional sphere (Extended Data Fig. 6a). The thicker and shorter horizontal black line inside the box is the location stereotypy, the average of all the values in that structure's block in the correlation matrix. Dots represent the raw data (one dot per correlation value; 1,000 randomly selected points are shown). The box extends from the first quartile (Q1) to the third quartile (Q3) of the data, with a line at the median. The whiskers extend from the box by 1.5x the interquartile range (IQR). Numbers of cells are in Supplementary Data 1. Colour bars along the bottom (x axis) indicate the cellular structure. Numbers above the colour bar indicate structures ranked by their stereotypy from greatest to least. The structures with the greatest location stereotypy were the nuclear envelope (lamin B1) and the plasma membrane (via CAAX domain of K-Ras, "*CAAX*"). These observations are effectively positive controls, because these two structures should be very similar to the cell and nuclear boundary shapes that were used as fixed points in the SHE interpolation. In decreasing order of stereotypy, the next highest were two nucleolar compartments, the Dense Fibrillar Component (DFC, via fibrillarin) and the Granular Component (GC, via nucleophosmin), followed by the ER (both Sec61 beta and SERCA) and microtubules. Structures with the least location stereotypy included those with a low number of discrete separated locations near the top or bottom of the cell such as centrioles (via centrin-2), desmosomes (desmoplakin), and matrix adhesions (paxillin) as well as structures with sparse, punctate locations such as cohesins (SMC-1A), endosomes (Rab-5A) and peroxisomes (PMP34). **b**. The process to create the Pearson correlation matrix for the 8-dimensional sphere (Extended Data Fig. 6a) was repeated for the reconstructed cell and nuclear shapes at each of the nine map points for each of the eight shape modes. Shown here are the resulting correlation matrices along Shape Mode 1. Each entry in this matrix represents the correlation between the cellular structure PILR of two cells. Thicker black lines within the matrix indicate the regions (blocks) of the matrix corresponding to cells with a tagged structure. The size of each dimension of each block corresponds to the number of cells. **c**. Heat maps of the difference in location stereotypy for each of the eight shape modes (SM). Each heat map represents a shape mode, each column represents the nine binned map points along that shape mode (Fig. 2b), and each row

represents a different cellular structure. Each heat map value corresponds to the stereotypy difference between the mean cell shape and the cell shape in the indicated shape mode bin for that cellular structure. **d**. Heat maps of the difference in location concordance between the mean cell shape and either, the −2σ and 2σ shape space map points (bottom and top triangles, respectively), for each of the eight shape modes. Numbers of cells are in Supplementary Data 1. Colour bars on the left of heat maps indicate the cellular structure. Additional concordance difference heat maps are available in Supplementary Data 1.

Source data

Extended Data Fig. 8 Statistical analysis of the variation in cell, nuclear and cellular structure sizes.

a. Heat map in four parts summarizing the results of a systematic, comparative analysis of the relationship between the volumes of the15 cellular structures validated for structural volume analysis and five cell and nuclear size metrics: the volume and surface area of the cell and the nucleus, and fifth, the cytoplasmic volume (the difference between cell and nuclear volumes), referred to as *cell vol*, *cell area*, *nuc vol*, *nuc area*, and *cyto vol*, respectively; Supplementary Methods). The number of cells in **a–k** are either all cells (n = 202,847) or per cellular structure (Extended Data Fig. 1d). The leftmost column (green heat map, *scaling rate*) indicates the percentage increase in structure volume given one doubling in cell volume over a well-represented volume range in the cell population (1160 to 2320 μm^3). For example, the volume of mitochondria increased by an average 84% (from 108 to 199 μm^3) for this doubling in cell volume (a doubling is an increase of 100%). The structures with the greatest relative scaling rates were the peroxisomes, followed closely by both nucleolar structures and then microtubules, all of which nearly doubled in structure volume with the doubling of cell volume. Simple linear regression was used to fit the data and to calculate the percent of the variation in cellular structure volumes that can be explained by each of the five cell and nuclear size metrics (next five columns in **a**, blue-red heat map, *explained variance*). The percent explained variance was substantially greater for some structures, such as mitochondria (54%) than for other structures, such as endosomes (2%). For

nuclear structures like the nucleolar DFC, more of the variance in their volumes could be explained by nuclear volume than by cell volume (77% vs. 68%, respectively). A multivariate model was applied to calculate the total percentage of the variance explained for each of these structures by the combination of all four cell and nuclear size metrics (centre single column, *all metrics*). At the lowest end were the centrioles, which are discrete structures that double in number during the cell cycle, but with a negligible volume increase. Centrioles should not get continuously bigger as cells grow and were thus invariant with all size metrics. At the highest end were the nuclear envelope and the plasma membrane, which, as expected, correlated well with nuclear and cell surface areas, respectively. Notably, the volumes of all three nuclear body structures (nucleolar DFC, GC, and speckles) had high explained variances. Cell and nuclear metrics show a large degree of collinearity, which makes it non-trivial to isolate the effect of one particular cell or nuclear metric on structure volume. The multivariate model was used to calculate the *unique* contributions of both cell size metrics, both nuclear size metrics, and each of the four metrics individually (last six columns, orange heat map, *unique explained variance*). For all five nucleus-related structures, the variance in structure volume was better explained by nuclear size metrics than by cellular size metrics. For the nuclear envelope, more of the variance was uniquely attributable to the nuclear surface area than nuclear volume; this anticipated result confirmed the validity of this approach. **b**. Scatterplot of nuclear vs. cell volumes for all cells, coloured based on an empirical density estimate. The green line is a running average and the grey line is the linear regression model, also used to calculate the scaling rate (see **a**). **c**. Line plots showing the scaling rate for three cellular structures (yellow line and numbers in top left corners). The regions filled in grey are the interquartile range (IQR) measured across cells that were binned in 10 cell volume bins. The xy axes to the far left are used to indicate the values of the tick marks in each of the three plots. **d–g**. Scatterplots and statistical measures as in (**b**), for mitochondria (**d**), endosomes (**e**), and nucleoli (DFC, **f** and **g**). **h**. Scatterplot of the relative volume scaling rate vs. the total percent explained variance for the 15 cellular structures. Error bars are 5-95% confidence intervals calculated via bootstrap (n = 100). Structures along top and right side are rank ordered. The structures with the lowest relative volume scaling rates were also the structures identified as having the lowest

explained variance (endosomes, centrioles). For most structures, however, relative scaling rates were at least 60%, consistent with the simple expectation that larger cells typically would also have larger organelles. Two structures whose volumes correlated most strongly with nuclear surface area (nuclear envelope, nuclear speckles) showed lower scaling rates. This was consistent with surface area generally scaling less quickly than volume. For example, doubling the size of a perfect sphere leads to only a 59% increase in its surface area. The peroxisomes stood out as exhibiting an unusual pattern of both a high relative volume scaling rate and great variability in peroxisome volume from cell to cell. **i**. Scatter plot of nuclear surface area vs. nuclear volume for all cells (blue points), cells with spherical nuclei (n = 19,927, brown points), perfect spheres (magenta dashed line) and linear and non-linear model fits on spherical cells or all cells (cyan and black as indicated; Supplementary Methods). The volume (V) and surface area (A) of a sphere don't scale linearly, instead $A \sim V^{2/3}$. However, on this dataset a linear model of nuclear volume explains as much variation in nuclear area as a model with the theoretically correct non-linear scaling factor. **j**. Scatterplot of explained variance for linear vs non-linear models for all cases in the heat map of explained variance in **a** (n = 190; Supplementary Methods). Median (across 100 bootstraps of the regression model; blue points) and 95% confidence interval (from 2.5% to 97.5% across the 100 bootstraps; red lines) are indicated. **k**. Heat map of percent explained variance between size-scaling metrics (rows) and shape modes (SM, columns). Correlations of structure volume to Shape Mode 5 likely occur due the moderate correlation between Shape Mode 5 (elongation) and cell surface area.

Source data

Extended Data Fig. 9 ALPs in shape-matched non-edge and edge cells.

a. Cell and nuclear SHE coefficients from a comparison dataset (e.g., edge cells; red dots) are transformed according to the SHE PCA of a baseline dataset (e.g. interphase cells; black + grey dots) resulting in the embedding of the comparison dataset cells into the baseline 8D shape space. Each cell in the comparison dataset is matched to its nearest neighbour in the shape

space that is also in the baseline dataset (lines connecting black and red dots), creating the shape-matched dataset. **b**. Average morphed cells for six cellular structures in shape-matched non-edge and edge cells. For five of these structures, the ALP is a redistribution of the structure towards the outer edge of the colony, while for adherens junctions (via beta-catenin) the ALP is a redistribution of junctions away from the colony edge. **c**. Dimensionality of PILRs of cells in the shape-matched dataset is first reduced to 32 via PCA (see Methods). LDA is then applied to these 32 PCs to find the axis of greatest separation (solid purple line) between the two groups of cells in the dataset (black and red dots). Data points are projected along the discriminant axis to determine the frequency of cells. **d**. Average morphed cells for actin bundles (via alpha-actinin-1) in non-edge and edge cells. **e**. PILR-LDA based reconstructions of actin bundles in average morphed cells at five positions (in σ units) along the LDA axis. Dotted lines correspond to the locations of the mean non-edge (black) and edge (red) cells in (**d**). **f**. Frequency of cells along the LDA axis within non-edge and edge cell populations. Dotted vertical lines indicate the means. **g**. Top view and side view 1 of three examples of each non-edge and edge cells along the LDA axis. Top row shows the original and bottom row the morphed visualizations for each of these cells. Images are average projections of the segmented structure. **h**. Frequency of cells along the LDA axis within non-edge and edge cell populations for the five structures in (**b**). Dotted vertical lines indicate the means. PILR-LDA based reconstructions of average morphed cells at five positions (in σ units) along the LDA axis for all 25 cellular structures as well as single-cell examples available in Supplementary Video 3. **i**. Heat maps of the differences in average location similarity (left), stereotypy (centre) and concordance (right) for the 25 cellular structures in shape-matched non-edge vs. edge cells (numbers of cells and heat map data in Supplementary Data 1).

Source data

Extended Data Fig. 10 Integrated intracellular reorganization in early mitosis (supporting figure).

a. We investigated two stages of early mitosis—prophase (m1) and early prometaphase (m2), when the condensing chromosomes still largely form

an aggregated, nuclear-like structure that could be biologically interpreted in the context of our cell and nuclear shape-based coordinate system. Due to the breakdown of the nucleus and the condensation of DNA in these early stages of mitosis, the outline of the DNA-dye-based segmentation was no longer appropriate for SHE-based parameterization. Instead, we replaced the nuclear segmentation of cells in both datasets with their convex-hull counterpart. **b**. Mean cell (magenta or purple) and nuclear (cyan or green) shape for all interphase cells (1^{st} column), cells in prophase (m1), shape-matched interphase 1 and m1 cells (i1 + m1), cells in early prometaphase (m2), and shape-matched interphase 2 and m2 cells (i2 + m2), respectively. **c**. Frequency of cells for the eight shape modes (SM) for all interphase (grey), i1 (black) and m1 (red) cells (top two rows), i2 (black) and m2 (red) cells (bottom two rows). **d**. Concordance heat maps for interphase cells in the two shape-matched interphase datasets (i1, i2) and their corresponding prophase (m1) and early prometaphase (m2) mitotic cells. **e**. Heat maps of the differences in concordance in early mitosis for i1–i2, i1–m1, m1–m2, and i2–m2 stages. **f**. Flagged significant concordance differences (black boxes) for each of the difference heat maps shown in (e). **g**. Average structure similarity heat maps for interphase cells in the two shape-matched interphase datasets (i1, i2) and their corresponding prophase (m1) and early prometaphase (m2) mitotic cells. Due to the low number of cells in mitosis for some structures, we did not quantitatively analyse differences in the average location similarities, although their qualitative results matched those based on the concordance values. Heat maps in Supplementary Data 1.

Source data

Extended Data Fig. 11 Summary of this study.

a. The Allen Cell Collection of high-quality gene-edited FP-tagged cell lines and the standardized microscopy imaging pipeline, combined with new tools for image analysis permitted us to create the WTC-11 hiPSC Single-Cell Image Dataset v1 of over 200,000 living cells and 25 cellular structures in 3D. **b**. We created two distinct conceptual coordinate systems to analyse our cells. The first maps the shape of an individual cell with respect to the total shape variation observed in the entire population via a

3D cell and nuclear shape space (via SHE). The second maps the location of every cellular structure within an individual cell (via the PILR). **c**. With these two coordinate systems we created an analysis framework to measure distinct aspects of integrated intracellular organization, including measurements of structure volume variations as well as the locations of cellular structures. This included the average locations both of individual structures and all pairs of structures (ALP and average structure similarities), as well as the variability in these locations (stereotypy and concordance). **d**. This suite of measurements was applied to our large baseline dataset of interphase cells and showed that integrated intracellular organization was very robust across the wide range of cell shapes in the normal interphase population. **e**. Two cell subpopulations stood out morphologically in the dataset: colony edge cells and mitotic cells, prompting us to assess their organization. To do this, we developed a process to match each individual cell in the chosen subpopulation with a "control" (interphase) cell of similar overall shape, and then used analysis of these shape-matched pairs to visualize and quantify the location phenotype of greatest difference between the two populations (via the PILR-LDA). **f**. First, we compared the intracellular organization of cells at the edges of hiPS cell colonies compared with shape-matched non-edge cells. We found that some structures showed a polarized location towards the colony edge but this change in location was not accompanied by any other changes in pairwise structure locations or variations, suggesting that while the locations changed, the variability and relationships among structures (average structure similarities, stereotypy, and concordance) i.e., the *"wiring"*, of the cell did not. **g**. In contrast, our second subpopulation comparison focused on early mitotic cells confirmed that they undergo a dramatic intracellular reorganization, in which not only the average locations of structures, but also their wiring, changed substantially. To assess these changes with a robust quantitative perspective, we developed new workflows to formally identify when significant changes in any of these measurements occurred in the first two early stages of mitosis, and then summarized and visualized these results in a way that could facilitate further data exploration and hypothesis generation. We found that all structures except those located at the cell periphery changed their average locations during early mitosis. Furthermore, all structures that changed location (other than the four for which stereotypy was statistically not

measurable) also changed in at least one other aspect of their organization (stereotypy, concordance, or both) during at least one of the two stages of early mitosis. Thus, structure location changes of cells in early mitosis, unlike in edge cells, were accompanied by changes in their wiring. This suggests that edge cells and early mitotic cells may represent distinct classes of cellular reorganization, perhaps related to the specific cellular processes underlying them. **h**. We performed a meta-analysis to investigate the association between distinct aspects of cell organization observed throughout this study. The results of this meta-analysis prompted us to suggest a possible hierarchy of dependencies as cells reorganize, as outlined in the **Discussion**. However, our observations also demonstrate that this simple proposed hierarchy among these distinct aspects of organization is not absolute. It is possible that these potential dependencies, or "*rules*" of cell organization, are general and apply to a range of genetic perturbations, differentiation, signalling factors, environmental signals, etc. It is also possible that there is a larger set of cell type or state-dependent organizational rules.

Extended Data Fig. 12 Statistical analysis for quality control of the WTC-11 hiPSC Single-Cell Image Dataset v1.

a. Box plots of principal component values for all cell lines together (first bin in dark green) and per tagged structure cell line, plotted in pipeline timeline order, the order that structure datasets were collected (total n = 175,147; n per structure in Supplementary Data 1). The box extends from the first quartile (Q1) to the third quartile (Q3) of the data, with a line at the median. The whiskers extend from the box by 1.5x the interquartile range (IQR and dots represent outliers beyond the IQR. The dashed horizontal line spanning the entire plot represents the median value for all cell lines together (first bin in dark green). The colours for each cell line refer to the pipeline workflow (see Methods for details). Triangles indicated structures for which the IQR does not overlap with the mean value for all cell lines. **b**. Left plots shows the distributions of cell height (top) and cell volume (bottom) for all cell lines together (first bin in dark green; n = 202,847) and per tagged structure cell line, plotted in pipeline timeline order (n per structure in Supplementary Data 1 and Extended Data Fig. 1d). Structure names in red indicate those structures imaged with an adjusted Matrigel

coating protocol towards the end of the pipeline timeline. The centre plots show a comparison of cell height (or volume, bottom) between actomyosin bundle-tagged cells (via non-muscle myosin IIB) in the main dataset (Pipeline 4.1; n = 6,223) and in a repeat dataset imaged with Pipeline 4.4 settings with the adjusted Matrigel coating protocol (n = 380). The right plots shows a comparison of cell height (or volume, bottom) between all cell lines imaged pre-Pipeline 4.4, during Pipeline 4.4 with original Matrigel coating and during Pipeline 4.4 with adjusted Matrigel coating. Percentages shown in the plot are the relative height reduction compared to the mean height of cell lines imaged pre-Pipeline 4.4. **c**. The top image diagrams circular mapping of imaged colonies (via the 12X overview images). Two cells are represented by two red dots within an FOV, represented by a rectangle. The FOV centre is at distance *d* from the closest edge of the colony. The two cells are then mapped into a unit circle that serves as a template to visualize the radial location of the two cells. The radial location is the FOV relative distance to the edge of the colony, $\ell = d/R_{eff}$, where R_{eff} represents the effective radius of the colony. The angular location of a cell (θ_1 and θ_2 for the two cells in the image) is independently drawn from a uniform distribution of angles in the range $[0,2\pi]$. Cells from the dataset that were associated with a colony size (see Methods) were grouped into four bins, each with similar number of cells, based on the area of the colony where they came from. The colony area range of each bin is 15k-230k μm^2, 230k-377k μm^2, 377-620k μm^2 and 620k-14,285k μm^2. Each point represents one cell within the colony area bin that was mapped into the unit circle. The unit circle was then rescaled to match the mean colony area for that bin. Points are colour-coded by their corresponding cell height. Listed above each circle is the mean colony area in that bin to which the unit circle is scaled. Below each circle are profile plots of cell height as a function of the radial distance for each of the cell (in black). The red curve represents the rolling average. Each row of circular colony mappings represents a different aggregation of the data based on the imaging mode: the first row is for all imaging modes (modes A, B and C; n = 104,269), the second row is for modes A and B only (n = 75,146) and third row is for mode C only (n = 29,123). **d**. Circular colony mappings as in (**c**) where points (cells) are now colour-coded by values of the shape modes. Circular colony mappings are shown for Shape Modes 1 and 2, and profile plots (as in **c**), for Shape Modes 3-8 (all imaging modes, n = 104,269). **e**. Scatter

plots on the far left show true values of cell height compared to cell height values predicted by random forest regression models (n = 95; see Methods) that include either all experimental variables (top plot) or all experimental variables except for the cell line identity (bottom plot). The error bars on the predicted values are obtained via bootstrapping (n = 100). The centre column shows box plots representing the feature importance for each of the two models as measured by the increase in the mean squared error (MSE) when all values of that corresponding feature are shuffled across samples. The box extends from the first quartile (Q1) to the third quartile (Q3) of the data, with a line at the median. The whiskers extend from the box by 1.5x the interquartile range (IQR and dots represent outliers beyond the IQR. The right top plot is the Pearson correlation matrix between five continuous experimental variables used in training the regression models. The bottom right plot is the Cramer’s V correlation matrix between six categorical experimental variables used in training the regression models. Variables with correlation above the significance threshold 0.3 are assumed to be highly correlated[53].

Source data

Supplementary information

Supplementary Methods

This document contains more information and details on experimental methods, quantification, and statistical analysis.

Reporting Summary

Supplementary Data 1

This DataFile contains 1) a summary of all of the numbers of FOVs, imaging days and cells for all analyses, 2) the heat map data for the average location similarities, stereotypy, and concordance, including difference heat maps, and 3) additional data on the comparative analysis of cellular structure volumes in edge and non-edge cells.

Supplementary Data 2

This DataFile contains a table of key resources, including materials, cell lines, and links for data and code availability.

Supplementary Video 1

Animated gif of the cell and nuclear shape space shown in Fig. 2d. 2D projections of 3D meshes obtained for each of the nine map point bins of each of the eight shape modes. All three views are shown for each mode, as indicated along the top. Human-interpretable names for these shape modes are indicated on the right. Mesh projections of the cell are in magenta and of the nucleus are in cyan. Each frame of the video shows a successive map point location along each shape mode, from -2σ to 2σ in steps of 0.5σ (σ = standard deviation).

Supplementary Video 2

3D visualization of 17 structures rendered simultaneously. This visualization illustrates the relative spatial relationships of the 17 structures as in Fig. 3c.

Supplementary Video 3

Overview panel of average morphed cells for 25 cellular structures reconstructed throughout the cell and nuclear shape space. Each row represents one of the 25 cellular structures, indicated by the colour bar on the far left. Each column represents Shape Modes 1-8, respectively. Each frame of the video represents one of the nine map points bins (Fig. 2b; in σ units, indicated at top left) along each shape mode and shows the resultant average morphed cells for the cells in that bin (numbers of cells in Supplementary Data 1). For every cell the top view and side view 1 are shown. Heat maps as in Extended Data Fig. 5. Scale bars are 5 µm. See https://www.allencell.org/viana-2022-supplementary-panels for improved viewing of the Video's content.

Supplementary Video 4

Overview panel video of 25 cellular structures in edge and shape-matched non-edge cells. a. Within each half, either moving left to right (non-edge cells) or right to left (edge cells), two sets of two individual cell examples of the structure segmentation, the result of morphing that cell's PILR into the combined non-edge + edge mean cell shape, and the resultant average morphed cells, visualized as in Extended Data Fig. 5. b. PILR-LDA based reconstructions of average morphed cells at five positions (in σ units; each position is a frame in the video, indicated by the arrowhead) along the non-edge to edge cell LDA axis. Histograms show frequency of cells along the LDA axis within non-edge (black) and edge (red) cell populations. Dotted vertical lines indicate the means. X indicates that LDA axis primarily identified technical differences in the PILR (Supplementary Methods). Heat maps as in Extended Data Fig. 5. Scale bars are 5 µm. See https://www.allencell.org/viana-2022-supplementary-panels for improved viewing of the Video's content.

Supplementary Video 5

Overview panel video of 25 cellular structures in early mitosis. a. Sets of two columns showing example(s) of individual cell structure segmentations and the resulting average morphed cells for all four early mitotic populations (i1, i2, m1, m2) in the appropriate mean mitotic cell shapes, visualized as in Extended Data Fig. 5. b. PILR-LDA based reconstructions of average morphed cells at five positions (in σ units; each position is a frame in the video, indicated by the arrowhead) along the i1 to m1 (left) and i2 to m2 (right) LDA axes. Histograms show the frequency of i1 and m1 cells or i2 and m2 cells along the LDA axis. Dotted vertical lines indicate the means. X indicates that LDA axis primarily identified technical differences in the PILR (Supplementary Methods). Heat maps as in Extended Data Fig. 5. Scale bars are 5 µm. See https://www.allencell.org/viana-2022-supplementary-panels for improved viewing of the Video's content.

Source data

Source Data Fig. 2

Source Data Fig. 3

Source Data Fig. 4

Source Data Fig. 5

Source Data Extended Data Fig. 3

Source Data Extended Data Fig. 4

Source Data Extended Data Fig. 6

Source Data Extended Data Fig. 7

Source Data Extended Data Fig. 8

Source Data Extended Data Fig. 9

Source Data Extended Data Fig. 10

Source Data Extended Data Fig. 12

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted

by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Viana, M.P., Chen, J., Knijnenburg, T.A. *et al.* Integrated intracellular organization and its variations in human iPS cells. *Nature* **613**, 345–354 (2023). https://doi.org/10.1038/s41586-022-05563-7

- Received: 13 January 2021
- Accepted: 15 November 2022
- Published: 04 January 2023
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05563-7

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

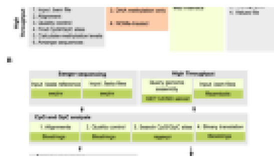
Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

Structural landscape inside cells mapped in detail

Nature Research Briefing 04 Jan 2023

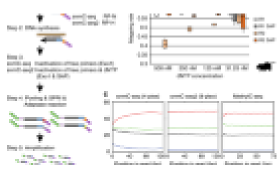
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05563-7



NOMePlot: analysis of DNA methylation and nucleosome occupancy at the single molecule

31 May 2019

Francisco Requena, Helena G. Asenjo, … David Landeira



Robust single-cell DNA methylome profiling with snmC-seq2

20 September 2018

Chongyuan Luo, Angeline Rivkin, … Joseph R. Ecker



Smart-RRBS for single-cell methylome and transcriptome analysis

09 July 2021

Hongcang Gu, Ayush T. Raman, … Andreas Gnirke



[iMETHYL: an integrative database of human DNA methylation, gene expression, and genomic variation]

29 March 2018

Shohei Komaki, Yuh Shiwa, … Atsushi Shimizu



[Highly scalable generation of DNA methylation profiles in single cells]

09 April 2018

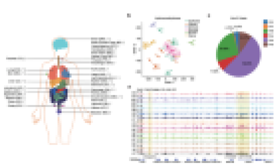
Ryan M Mulqueen, Dmitry Pokholok, … Andrew C Adey



[Targeted in situ genome-wide profiling with high efficiency for low cell numbers]

12 April 2018

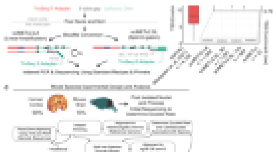
Peter J Skene, Jorja G Henikoff & Steven Henikoff



[Tissue-specific 5-hydroxymethylcytosine landscape of the human genome]

12 July 2021

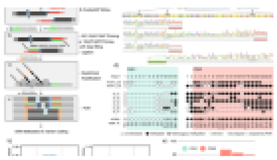
Bo He, Chao Zhang, … Chengqi Yi



[High-throughput robust single-cell DNA methylation profiling with sciMETv2]

09 December 2022

Ruth V. Nichols, Brendan L. O'Connell, … Andrew C. Adey



[Bisulfite-free epigenomics and genomics of single cells through methylation-sensitive restriction]

01 February 2021

Christoph Niemöller, Julius Wehrle, … Heiko Becker

- Article
- Open Access
- Published: 04 January 2023

A DNA methylation atlas of normal human cell types

- Netanel Loyfer,
- Judith Magenheim,

- Ayelet Peretz,
- Gordon Cann,
- Joerg Bredno,
- Agnes Klochendler,
- Ilana Fox-Fisher,
- Sapir Shabi-Porat,
- Merav Hecht,
- Tsuria Pelet,
- Joshua Moss,
- Zeina Drawshy,
- Hamed Amini,
- Patriss Moradi,
- Sudharani Nagaraju,
- Dvora Bauman,
- David Shveiky,
- Shay Porat,
- Uri Dior,
- Gurion Rivkin,
- Omer Or,
- Nir Hirshoren,
- Einat Carmon,
- Alon Pikarsky,
- Abed Khalaileh,
- Gideon Zamir,
- Ronit Grinbaum,
- Machmud Abu Gazala,
- Ido Mizrahi,
- Noam Shussman,
- Amit Korach,
- Ori Wald,
- Uzi Izhar,
- Eldad Erez,
- Vladimir Yutkin,
- Yaacov Samet,
- Devorah Rotnemer Golinkin,
- Kirsty L. Spalding,
- Henrik Druid,

- Peter Arner,
- A. M. James Shapiro,
- Markus Grompe,
- Alex Aravanis,
- Oliver Venn,
- Arash Jamshidi,
- Ruth Shemer,
- Yuval Dor ,
- Benjamin Glaser &
- …
- Tommy Kaplan

Show authors

Nature **volume 613**, pages 355–364 (2023)

- 20k Accesses
- 205 Altmetric
- Metrics details

Abstract

DNA methylation is a fundamental epigenetic mark that governs gene expression and chromatin organization, thus providing a window into cellular identity and developmental processes[1]. Current datasets typically include only a fraction of methylation sites and are often based either on cell lines that underwent massive changes in culture or on tissues containing unspecified mixtures of cells[2,3,4,5]. Here we describe a human methylome atlas, based on deep whole-genome bisulfite sequencing, allowing fragment-level analysis across thousands of unique markers for 39 cell types sorted from 205 healthy tissue samples. Replicates of the same cell type are more than 99.5% identical, demonstrating the robustness of cell identity programmes to environmental perturbation. Unsupervised clustering of the atlas recapitulates key elements of tissue ontogeny and identifies methylation patterns retained since embryonic development. Loci uniquely

unmethylated in an individual cell type often reside in transcriptional enhancers and contain DNA binding sites for tissue-specific transcriptional regulators. Uniquely hypermethylated loci are rare and are enriched for CpG islands, Polycomb targets and CTCF binding sites, suggesting a new role in shaping cell-type-specific chromatin looping. The atlas provides an essential resource for study of gene regulation and disease-associated genetic variants, and a wealth of potential tissue-specific biomarkers for use in liquid biopsies.

Main

Understanding how the same DNA sequence is interpreted differently in different cell types is a fundamental challenge of biology. Gene expression, DNA accessibility and chromatin packaging are well-established essential determinants of cellular phenotype. Underneath these lies DNA methylation, a stable epigenetic mark that underpins the lifelong maintenance of cellular identity.

Available human DNA methylation datasets suffer from major limitations. Multiple studies that have characterized methylomes of embryonic development, differentiation, cancer or other settings[6,7,8,9] have relied on the Illumina BeadChip platforms, which are limited to a predefined subset of 450,000 or 860,000 CpG methylation sites, representing just 3% of around 30 million CpG sites in the human genome[10]. In addition, by measuring each CpG site independently, such assays overlook coordinated patterns of DNA methylation occurring in blocks, the critical functional units of DNA methylation[11,12].

Most DNA methylation analyses interrogated primarily bulk tissue, thus precluding the study of minority cell types such as tissue-resident immune cells, fibroblasts or endothelial cells, whereas others analysed cultured cells, which may contain nonphysiological methylation patterns introduced in vitro[13]. As a partial solution, recent studies used single-cell RNA sequencing data from whole tissues to identify marker genes expressed in specific cell types, then identified specific CpGs whose methylation is anticorrelated with expression. These could be used on array-based methylomes to deconvolute bulk tissue and assess cell type composition or sample purity[14,15], but might

be insufficiently accurate for identification of rare cellular contributions in liquid biopsies. Some studies of the human methylome did analyse isolated primary cells using whole-genome bisulfite sequencing (WGBS), but their scope was limited[2,4,5].

To overcome these limitations and to accurately characterize the human cell methylome, we performed deep genome-wide sequencing with paired-end, 150 base pair (bp)-long reads at an average sequencing depth of 30× (6.62× or greater) on fluorescent activated cell sorter (FACS)-purified populations of 39 human cell type groups obtained from freshly dissociated adult healthy tissues. We coalesced methylation patterns across the entire genome into blocks of homogeneously methylated CpG sites and used these to study variation in methylation patterns across cell types. Here we identify and characterize genomic regions that are uniquely methylated in a tissue or cell-type-specific manner, provide vignettes of their possible biological function and introduce a fragment-level deconvolution algorithm with applications such as clinical diagnosis based on circulating cell-free DNA methylation.

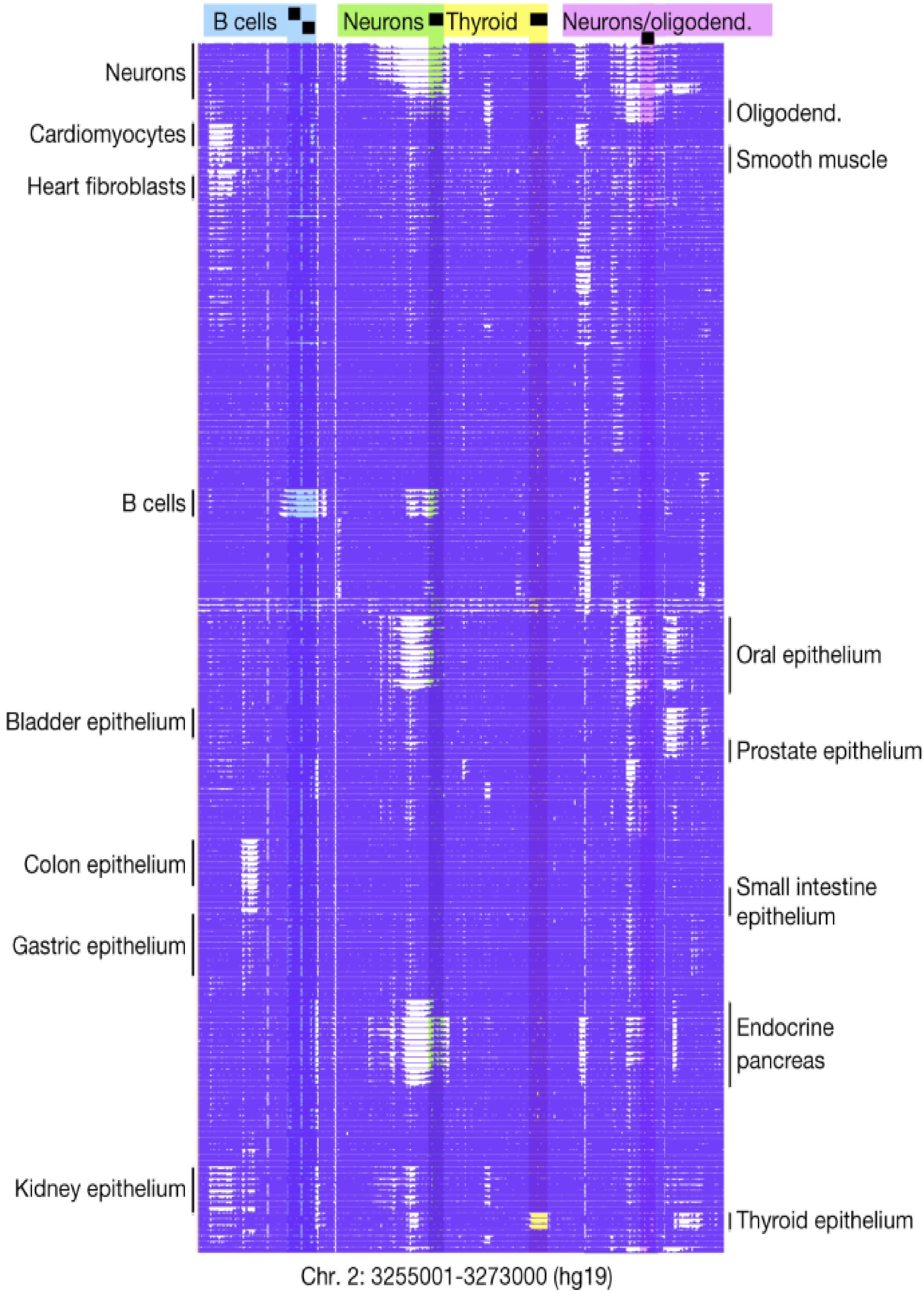
Methylation atlas of human cell types

To portray genome-wide DNA methylation across a variety of cell types, we performed WGBS (150-bp-long paired-end reads to a mean depth of at least 30×) on 205 samples representing 77 primary cell types from 137 consenting donors. These were carefully sorted and mapped to the human genome (hg19, hg38). Average sample purity (that is, proportion of material from desired cell type) was over 90% as determined by flow cytometry, gene expression and DNA methylation analysis. Several samples showed lower purity (for example, colon fibroblasts 78%, smooth muscle cells (SMC) 82%, endothelial cells 86% or adipocytes 87%). Detailed descriptions of sample isolation and purity estimations, as well as sample information, are provided in Supplementary Table 1, Supplementary Figs. 1–3 and Supplementary Information.

The cell types analysed (Fig. 1) represent most major human cell types, allowing a composite view of physiological systems (for example, gastrointestinal tract, haematopoietic cells and pancreas), as well as a

comparison of similar cell types in different environments (for example, tissue-resident macrophages).

Fig. 1: Methylation atlas of the adult human body.



B cells
Neurons
Thyroid
Neurons/oligodend.
Neurons
Cardiomyocytes
Heart fibroblasts
B cells
Bladder epithelium
Colon epithelium
Gastric epithelium
Kidney epithelium
Oligodend.
Smooth muscle
Oral epithelium
Prostate epithelium
Small intestine epithelium
Endocrine pancreas
Thyroid epithelium
Chr. 2: 3255001-3273000 (hg19)

DNA methylation patterns of 205 methylomes (rows) across 344 CpG sites (columns) are demonstrated in a 18 kb region. Highlighted are regions unmethylated specifically in B cells (blue), neurons (green), thyroid epithelium (yellow) and neurons/oligodendrocytes (oliogodend.) (pink).

The 205 methylomes show great similarities between replicates with distinctive changes between cell types in a block-like manner, as shown in Fig. 1. We sought to identify genomic regions differentially methylated in specific cell types to shed light on cell-type-specific biological processes, define cell identity and facilitate development of methylation biomarkers to identify the cellular origin of circulating cfDNA fragments[1,11,12,16,17,18,19,20,21].

We developed wgbstools, a computational machine learning suite, to represent, compress, visualize and analyse WGBS data (https://github.com/nloyfer/wgbs_tools). We segmented the genome into 7,104,162 nonoverlapping continuous blocks by identification of change points in DNA methylation patterns across multiple conditions. Each block spans highly correlated CpG sites similarly methylated in each sample but that may covary across cell types (Supplementary Information). We retained 2,783,421 methylation blocks of at least three CpGs with an average length of 544 bp (interquartile range (IQR) = 565 bp) and eight CpGs (IQR = 5 CpGs). Robust analysis of these compact genomic units is more straightforward than individual CpG sites and because of the regional nature of methylation can be viewed as the biological ‘atoms’ of human DNA methylation[12].

Interindividual variation in methylation

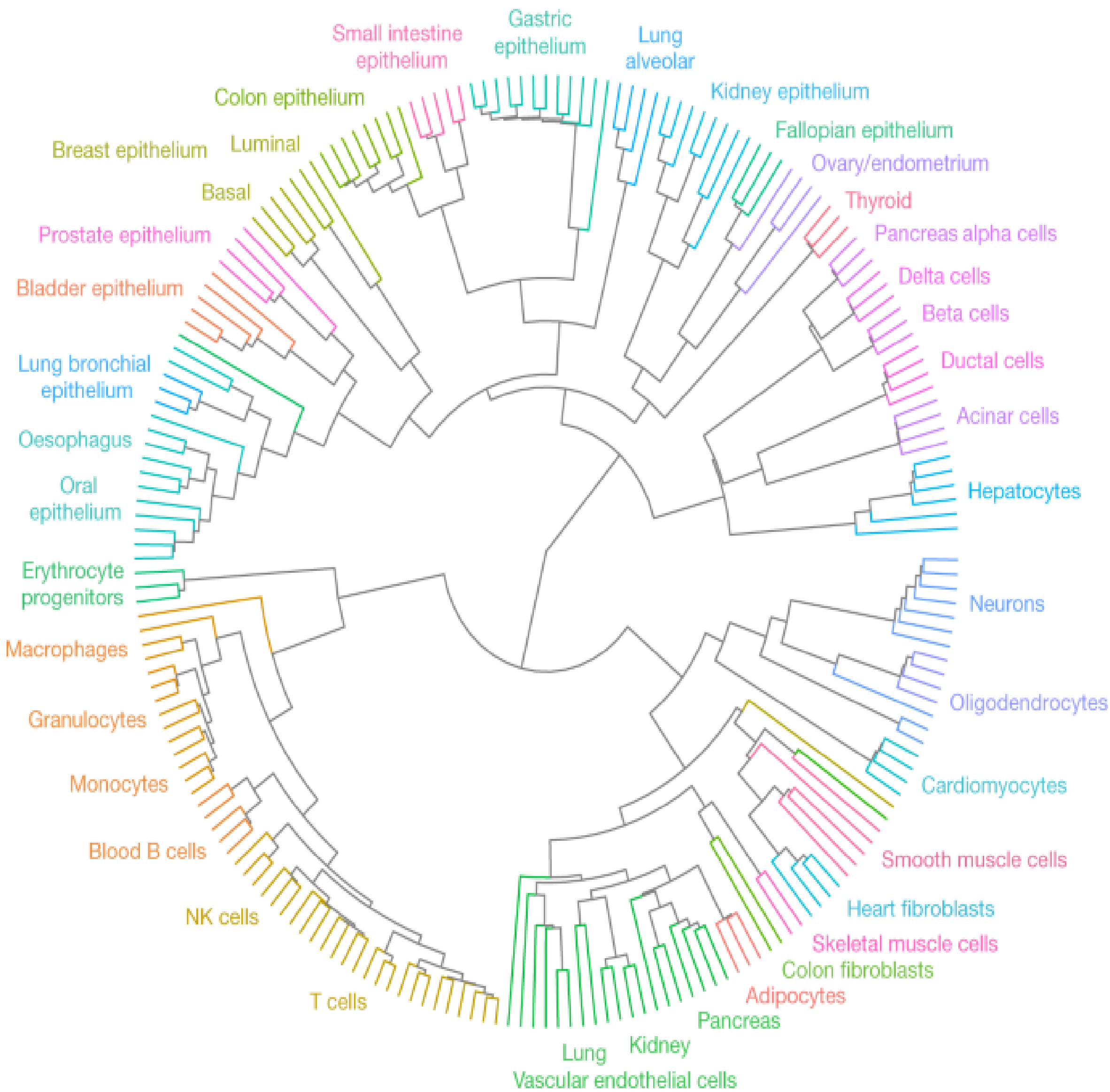
Methylation patterns were extremely robust across different individuals. For most cell types, 0.5% or less of blocks showed a difference of 50% or more across different donors compared with 4.9% among samples of different cell types (Extended Data Fig. 4). This high similarity in DNA methylation across donors is on a par with the estimated interindividual variability of genomic sequence[22]. Whereas the definition of 50% is somewhat arbitrary, other thresholds (35–50%) show a similar trend, with 0.5% or less variable blocks. Similar interindividual variation was observed in replicates obtained

from different laboratories (Supplementary Table 1). Strikingly, for cell types with $n \geq 3$ biological replicates, 195 of 197 samples (99%) showed the highest similarity to another replicate (rather than to another cell type from the same donor). These results demonstrate the reproducibility of preparations but also, in agreement with previous studies[6], highlight the fundamental biological phenomenon that DNA methylation is primarily determined by cell lineage and cell-type-specific programmes rather than by genetic or environmental factors.

Methylation records developmental history

Whereas DNA methylation patterns reflect the functional identity of a cell, they could also be used to track its developmental history. To identify patterns shared by the progeny of early progenitors, we calculated average methylation within blocks of at least four CpGs and selected those showing the highest variability across all samples (21,000 blocks, top 1%; Supplementary Table 2). We then clustered all 205 methylomes using an unsupervised agglomerative algorithm that iteratively identifies and connects the two closest samples regardless of their labelling[23]. This analysis systematically grouped biological samples of the same cell type (Fig. 2), similar to array-based clustering of purified human blood cells[6]. This supports the reproducibility of cell isolation and suggests that three or four replicates of each normal cell type are sufficient to infer its methylation patterns for practical applications such as biomarker identification.

Fig. 2: Unsupervised agglomerative clustering reflects human developmental lineage of healthy cell types.



Cell types are indicated by edge colours.

Strikingly, the resulting fanning diagram recapitulates key elements of lineage relationships among human tissues. For example, pancreatic islet cell types (alpha, beta and delta), which originate from the same embryonic endocrine progenitor[24], densely cluster together. Consistent with methylomes reflecting lineage rather than function, islet cells further cluster with pancreatic duct and acinar cells, and then with hepatocytes, with whom they share endodermal origins. Conversely, endoderm-derived islet cells do not cluster with ectoderm-derived neurons[25] despite common tissue-specific gene regulation and exocytosis machinery[26].

Additional examples include the clustering of gastric, small intestine and colon epithelial cells; the clustering of all blood cell types; and the clustering of multiple mesoderm-derived cell types including vascular endothelial cells, adipocytes and skeletal muscle. Interestingly, lung bronchial epithelium clustered with oesophagus and oral epithelium whereas lung alveolar epithelium clustered with intestinal epithelium, consistent with evidence of early developmental origins of the alveolar cell lineage[27].

Some methylation patterns were common to lineages that formed during early developmental stages. For example, 892 regions were unmethylated in epithelial cells derived from early endodermal derivatives and methylated in mesoderm- and ectoderm-derived cells (Methods). We suggest that these were demethylated in the endoderm germ layer, with derived cell types retaining these patterns decades later (Extended Data Fig. 5a). Because endoderm derivatives do not share common function or gene expression, this provides yet another example of methylation patterns as a stable lineage mark.

Finally, we applied the same segmentation and clustering approach to a published methylation atlas from the Roadmap Epigenomics project[4]. The algorithm did not group related cell types, and often clustered samples based on donor identity. This further emphasizes the importance of careful purification of homogeneous cell types, avoiding mixed cell populations (Extended Data Fig. 5b).

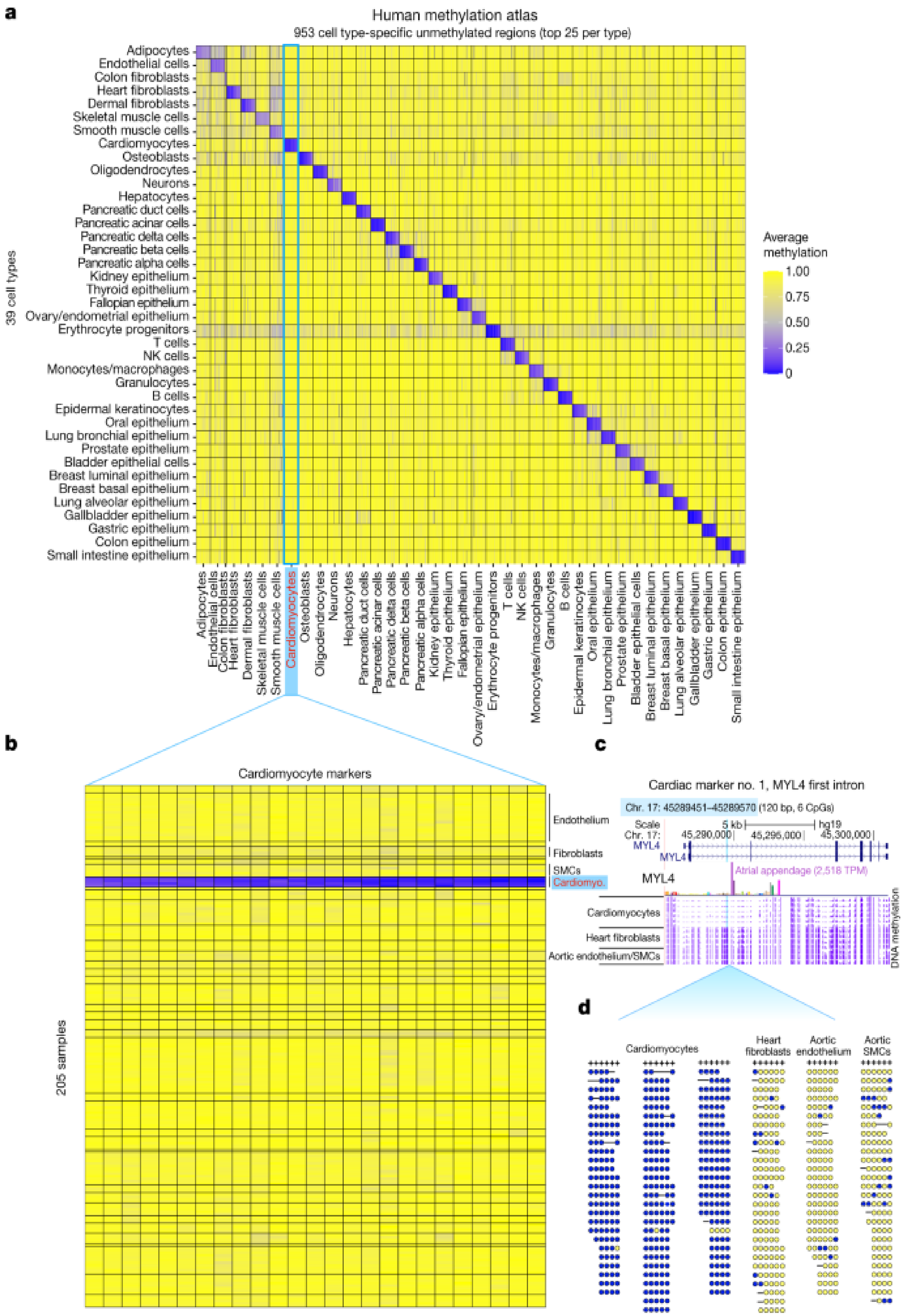
Cell-type-specific methylation markers

We next turned to study genomic regions differentially methylated in a cell-type-specific manner. We organized the 205 samples into 39 groups of specific cell types, including blood cell types (B, T, natural killer (NK), granulocytes, monocytes and tissue-resident macrophages), breast epithelium (basal and luminal), lung epithelium (alveolar and bronchial), pancreatic endocrine (alpha, beta and delta) and exocrine (acinar and duct) cells, vascular endothelial cells from various sources, cardiomyocytes and cardiac fibroblasts and more. We also defined 12 supergroups in which related cell types were grouped, including muscle cells, gastrointestinal epithelium, pancreas and more (Supplementary Table 3).

We then focused on differentially methylated blocks comprising five or more CpGs that are unmethylated in one group of cell types but methylated in all other samples, or vice versa. Intriguingly, almost all regions (97%) were unmethylated in one cell type and methylated in all others. We then sorted these differential regions by absolute difference in methylation in target cell type versus all other samples (Methods and Supplementary Information).

The top 25 differentially unmethylated regions for each cell type comprise a human cell-type-specific methylation atlas of 1,246 markers (Fig. 3 and Supplementary Table 4). These regions are uniquely unmethylated in particular cell types (average methylation 13%) and methylated in all other samples (average methylation 91%), and can serve as sensitive biomarkers for quantification of the presence of DNA from a specific cell type in a mixture. The markers include 953 cell-type-specific unmethylated loci, as well an additional 293 loci that are unmethylated in few related cell types. A fragment-level analysis further shows that the vast majority of DNA fragments at these regions are unmethylated in the target cell type compared with almost none in all other cell types (Extended Data Fig. 6). The atlas has various applications, including the analysis of circulating cell-free DNA fragments[18,19,20,21,28,29,30]. Importantly, only about 1% of cell type-specific markers are covered by reduced representation bisulfite sequencing (RRBS), 4–8% by methyl-sequencing hybrid capture panels and 14–24% are represented in single-CpG 450K/EPIC arrays[10], emphasizing the benefits of whole-genome sequencing for exhaustive identification of biomarkers.

Fig. 3: A human methylation atlas of 205 samples across 39 cell type groups.


a
Human methylation atlas
953 cell type-specific unmethylated regions (top 25 per type)
39 cell types
Adipocytes
Endothelial cells
Colon fibroblasts
Heart fibroblasts
Dermal fibroblasts
Skeletal muscle cells
Smooth muscle cells
Cardiomyocytes
Osteoblasts
Oligodendrocytes
Neurons
Hepatocytes
Pancreatic duct cells
Pancreatic acinar cells
Pancreatic delta cells
Pancreatic beta cells
Pancreatic alpha cells
Kidney epithelium
Thyroid epithelium
Fallopian epithelium
Ovary/endometrial epithelium
Erythrocyte progenitors
T cells
NK cells
Monocytes/macrophages
Granulocytes
B cells
Epidermal keratinocytes
Oral epithelium
Lung bronchial epithelium
Prostate epithelium
Bladder epithelial cells
Breast luminal epithelium
Breast basal epithelium
Lung alveolar epithelium
Gallbladder epithelium
Gastric epithelium
Colon epithelium
Small intestine epithelium
Average methylation
1.00
0.75
0.50
0.25
0
b
Cardiomyocyte markers
205 samples
Endothelium
Fibroblasts
SMCs
Cardiomyo.
c
Cardiac marker no. 1, MYL4 first intron
Chr. 17: 45289451–45289570 (120 bp, 6 CpGs)
Scale
Chr. 17:
5 kb
hg19
45,290,000
45,295,000
45,300,000
MYL4
MYL4
MYL4
Atrial appendage (2,518 TPM)
Cardiomyocytes
Heart fibroblasts
Aortic endothelium/SMCs
DNA methylation
d
Cardiomyocytes
Heart fibroblasts
Aortic endothelium
Aortic SMCs

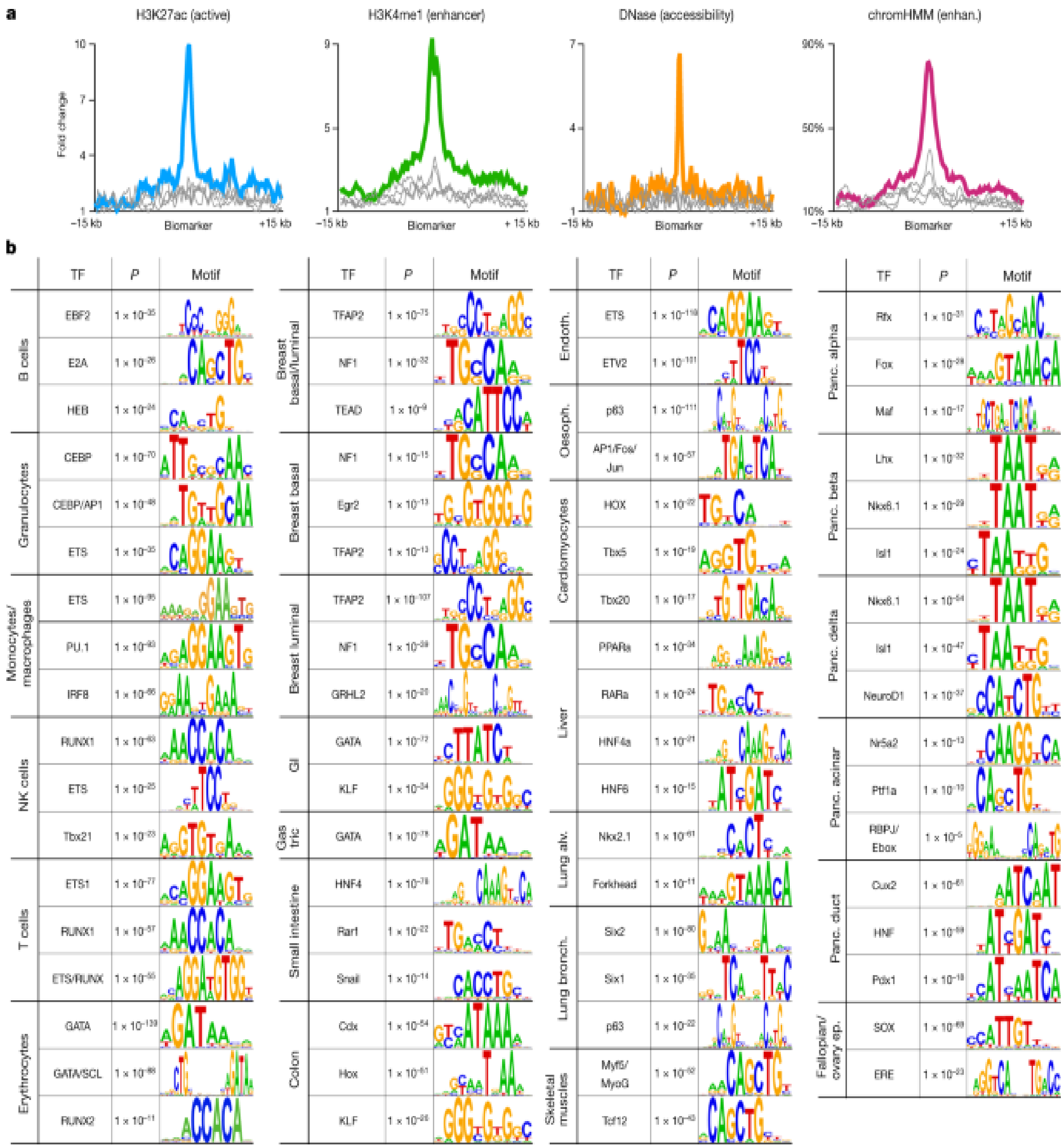
a, A total of 953 genomic regions unmethylated in a cell-type-specific manner. Each cell in the plot marks the average methylation of one genomic region (column) at each of 39 cell types (rows). Up to 25 regions are shown per cell type, with a mean length of 356 bp (nine CpGs) per region. **b**, The top 25 cardiomyocyte regions. For each region we plot the average methylation of each CpG site (columns) across all 205 samples in the atlas, grouped into 39 cell types as before. **c**, A locus specifically unmethylated in cardiomyocytes. This marker (highlighted in light blue) is 120 bp in length (six CpGs) and is located in the first intron of MYL4, a heart-specific gene (transcripts per million (TPM) expression of 2,518 in atrial appendage, GTEx inset). Genomic snapshot depicts average methylation (purple tracks) across six cardiomyocyte samples, four cardiac fibroblast samples and three aorta samples (two endothelial and one SMC). **d**, Visualization of bisulfite-converted fragments from three cardiomyocyte samples, one cardiac fibroblast sample and two aorta samples (endothelium and SMC). Shown are reads mapped to chr. 17: 45289451–45289570 (hg19), with at least three covered CpGs. Yellow and blue dots denote methylated and unmethylated CpG sites, respectively.

Human cell-type-specific regulatory maps

We next turned to characterize these sets of cell-type-specific differentially unmethylated regions. For this we identified the top 250 unmethylated markers for each cell type (Supplementary Table 4b) and used GREAT to identify those genes adjacent to each group of markers, and to test their enrichment for various gene-set annotations[31]. Genes adjacent to loci uniquely unmethylated in a given cell type typically reflect the functional identity of that cell type. For example, genes near B cell markers were enriched for B cell morphology, differentiation, IgM levels and lymphopoiesis; NK cell markers were associated with NK cell-mediated cytotoxicity, the haematopoietic system, cytotoxicity and lymphocyte physiology; Fallopian tube markers were enriched for egg coat and perivitelline space; and cardiomyocyte markers for cardiac relaxation, systolic pressure, muscle development and hypertrophy (Supplementary Table 5).

We then analysed the DNA accessibility and chromatin packaging of cell-type-specific markers as defined by assay for transposase-accessible chromatin using sequencing (ATAC–seq), DNase I hypersensitive site sequencing (DNaseI–seq)[4,32] and histone marks indicative of active promoters and enhancers[4]. The top 250 unmethylated markers for monocytes and macrophages are highly accessible and characterized by H3K27ac and H3K4me1 in monocytes, whereas markers of other cell types show no enrichment in monocytes (Fig. 4a), with similar results for markers of other cell types (Extended Data Fig. 7). We also show strong coordinated enrichment of chromHMM enhancer annotations at cell-type-specific markers[33] (Fig. 4a). These findings are consistent with previous studies that have associated tissue-specific demethylation with gene enhancers[1,34].

Fig. 4: Cell-type-specific markers as putative enhancers.



a, Average ChIP–seq signal for the active regulatory marker H3K27ac, enhancer marker H3K4me1, DNA accessibility and chromHMM enhancer annotations for the top 250 cell-type-specific unmethylated markers for monocytes/macrophages. Average signal for the top 250 markers of other blood cell types (granulocytes and B, T and NK cells) shown as grey lines, for comparison. **b**, Cell-type-specific markers are enriched for regulatory motifs. Shown are the top TF binding site motifs, enriched among the top 1,000 differentially unmethylated regions per cell type, using HOMER motif analysis. Motifs similar to previous (more significant) hits not included.

Shown are HOMER binomial *P* values. Alv., alveolar; Bronch., bronchial; Endoth., endothelium; Ep., epithelium; Oesoph., oesophagus; Panc., pancreas.

To further assess the biological importance of cell-type-specific unmethylated regions, we studied their association with transcription factors (TFs) that could either affect DNA methylation or bind DNA in a cell-type-specific manner, depending on methylation and chromatin[35,36,37,38]. We identified the top 1,000 unmethylated markers per cell type (Supplementary Table 4c) and performed motif analysis using HOMER[39] to calculate the enrichment of known TF binding motifs (Supplementary Table 6a). For most cell types the top motifs included master regulators and key TFs (Fig. 4b). For example, B cells are enriched for Ebf2/HEB/E2A, granulocytes for CEBP/AP1/ETS and T cells for ETS/RUNX. This association between cell-type-specific unmethylated regions and TF binding motifs can identify new gene regulatory circuits and expose distal enhancers active in specific cell types.

We aimed to identify the target genes of putative enhancers marked by cell-type-specific demethylation. Top markers frequently fall within intronic regions and are likely to regulate these genes (for example, glucagon in pancreatic alpha cells, NPPA and MYL4 in cardiomyocytes and MBP in oligodendrocytes; Supplementary Table 7), or proximally to probable targets (for example, a beta cell marker 5 kb from the insulin gene). Other markers are further apart from their target genes. We devised a computational algorithm to identify genes in the proximity of cell-type-specific markers showing increased gene expression levels under matching conditions (Methods). This highlighted hallmark genes for many cell types and suggested putative targets for many of the top 25 unmethylated markers for each cell type. For example, hepatocyte markers were associated with APOE, APOC1, APOC2 and the glucagon receptor. Similarly, cardiomyocyte markers were associated with NPPA, NPPB and myosin genes; and pancreatic islet markers with insulin and glucagon genes (Supplementary Table 7). These findings further support the principle that loci specifically unmethylated in a given cell type are probably enhancers positively regulating genes expressed in this cell type, often controlling adjacent genes. We note, however, that genes adjacent to a locus specifically

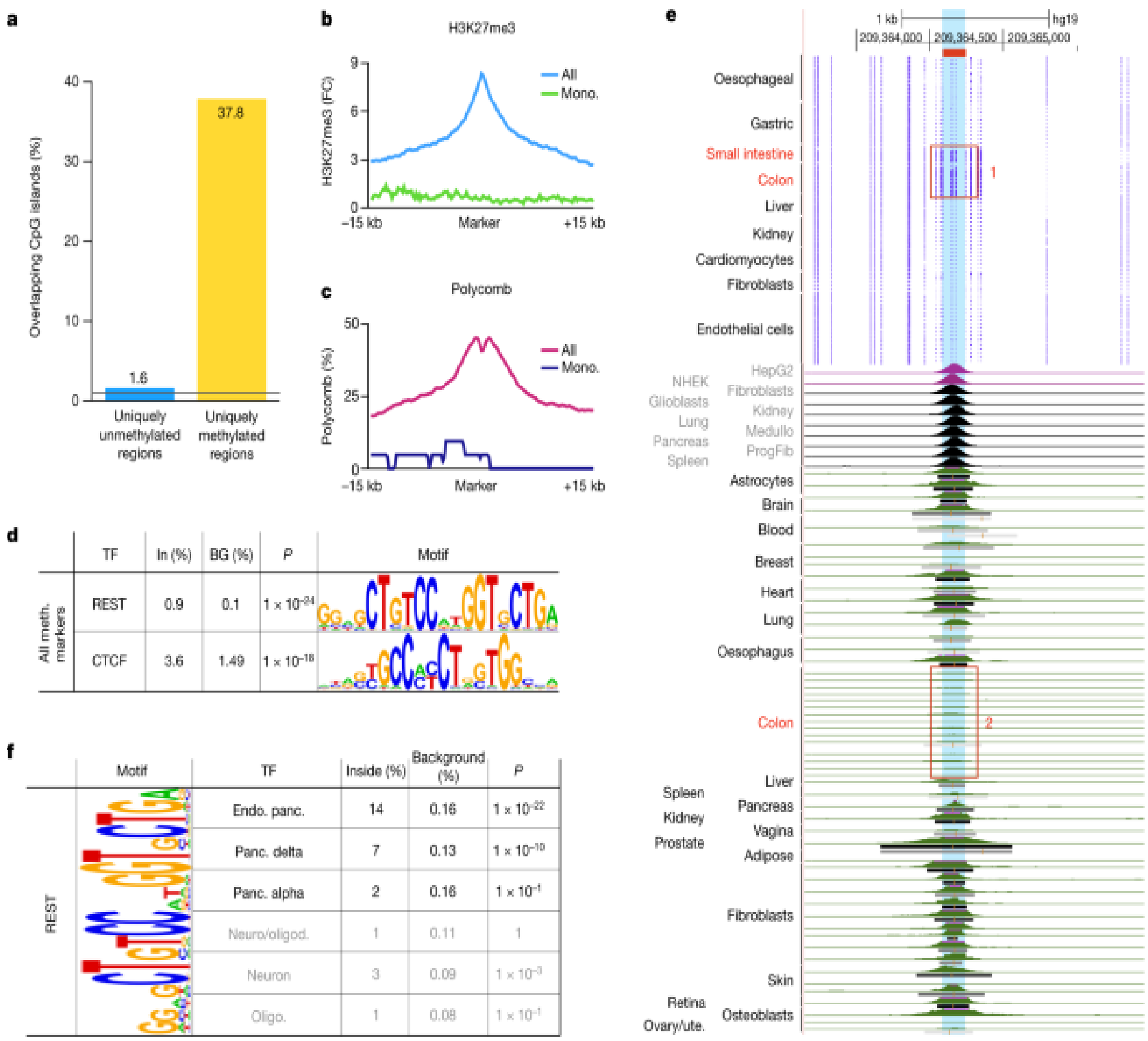
unmethylated in a given cell type are often broadly expressed beyond this cell type (Discussion).

To generate a catalogue of putative regulatory regions in each cell type we applied a fragment-level analysis across all samples from each cell type, independently of other cell types. We scanned the entire genome and identified genomic regions in which at least 85% of DNA fragments with at least four CpGs are unmethylated (Methods). This identified a set of unmethylated genomic regions in each of the 39 cell type groups analysed, including 36,111 regions on average (Supplementary Dataset 1). These regions were then annotated for genomic features, showing that 56% on average overlapped CpG islands, 46% were near promoter regions and 44% overlapped CTCF binding sites, thus highlighting the regulatory and structural roles of unmethylated loci. When available, we crossed these regions with chromatin immunoprecipitation sequencing (ChIP–seq) peaks from ENCODE[5] and Roadmap Epigenomics[4] under matching conditions, including H3K4me3, H3K27ac, H3K4me1, H3K27me3, CTCF and ATAC–seq, and generated a cell-type-specific catalogue of putative enhancer regions comprising unmethylated regions that overlap H3K27ac, but not H3K4me3, peaks (Supplementary Dataset 2). Motif analysis of these regions identified key TFs in each cell type, similar to those shown in Fig. 4 (Supplementary Table 6b,c).

Cell-type-specific hypermethylated loci

We studied those genomic regions methylated in one cell type but unmethylated elsewhere in the human body. These are enriched for CpG islands (38% of methylated regions compared with 1.7–2.7% of cell-type-specific unmethylated regions), and are marked by H3K27me3 and Polycomb in other cell types (Fig. 5a–c), as previously reported for cancer and developmental processes[40,41]. These cell-type-specific hypermethylated regions were generally less significant for motif enrichment (compared with uniquely unmethylated regions). Intriguingly, only around 3% of the total set of cell-type-specific differentially methylated regions are hypermethylated.

Fig. 5: Cell-type-specific hypermethylated regions are enriched for CpG islands, Polycomb targets and CTCF and REST/NSRF.



a, Thirty-eight per cent of the top cell type-specific hypermethylated markers (1,363 of 3,613, binomial $P < 1 \times 10^{-100}$) overap CpG islands. By comparison, 1.6% of cell-type-specific hypomethylated regions (189 of 11,714) overlap CpG islands, comprising less than 0.9% of the genome (black line). **b**, These regions are typically enriched for H3K27me3 in other cell types. Shown are average H3K27me3 signals in monocytes and macrophages near all cell-type-specific hypermethylated regions (blue) or near monocyte/macrophage-specific hypermethylated regions (mono; green). **c**, Similar plots for Polycomb annotations in monocytes and macrophages (chromHMM), for all or monocyte/macrophage-specific markers. **d**, Motif analysis of cell-type-specific hypermethylated regions (top 100 per cell type) identifies known CTCF and REST/NSRF motifs. HOMER binomial P values are shown. **e**, Analysis of ChIP–seq data for one such site

(chr. 1: 209364093–209364250, highlighted in blue, hg19), specifically methylated in the small intestine and colon epithelium (red box 1) and unmethylated elsewhere. As shown below, this site is bound in multiple cell types and tissues but is mostly unbound in stomach and colon epithelium in vivo (red box 2). **f**, REST/NSRF motif is present within 14% of the top 100 cell type-specific hypermethylated regions in the endocrine pancreas, 7% of top delta cell markers and 2% of top alpha cell markers, compared with approximately 0.1% in background sequences, in accordance with REST target expression in the endocrine pancreas. HOMER binomial *P* values are shown. Alv., alveolar; bronch., bronchial; Endo. panc., endocrine pancreas; Ep., epithelium; Oesoph., oesophagus; Oligo, oligodendrocytes; Panc., pancreas; Ute., uterus.

After pooling all cell-type-specific hypermethylated regions, we identified strong enrichment for target sequences of the chromatin regulator CTCF ($P \leq 1 \times 10^{-18}$; Fig. 5d). This suggests that DNA methylation of CTCF binding sites could act as a tissue-specific regulatory switch to modulate its binding, potentially affecting tissue-specific three-dimensional genomic organization[35,42,43]. To test this idea we compared patterns of DNA methylation at CTCF sites with genome-wide CTCF protein binding in specific tissues. Figure 5e shows the methylation pattern and published in vivo CTCF occupancy at one locus, which is methylated specifically in the colon and intestine. Consistent with DNA methylation preventing CTCF binding, ChIP data show selective absence of CTCF binding at this locus in the colon. In addition, loci methylated in specific cell types were enriched for targets of the transcriptional repressor of neural genes, RE1-silencing TF/neuron-restrictive silencer factor (REST/NRSF) ($P \leq 1 \times 10^{-24}$), and this was seen most prominently in the methylome of pancreatic islet cells (Fig. 5f). Whereas DNA methylation has not been shown to affect the binding or activity of REST, this finding raises the intriguing possibility that methylation of REST targets in islets could permit endocrine differentiation independently of REST repression.

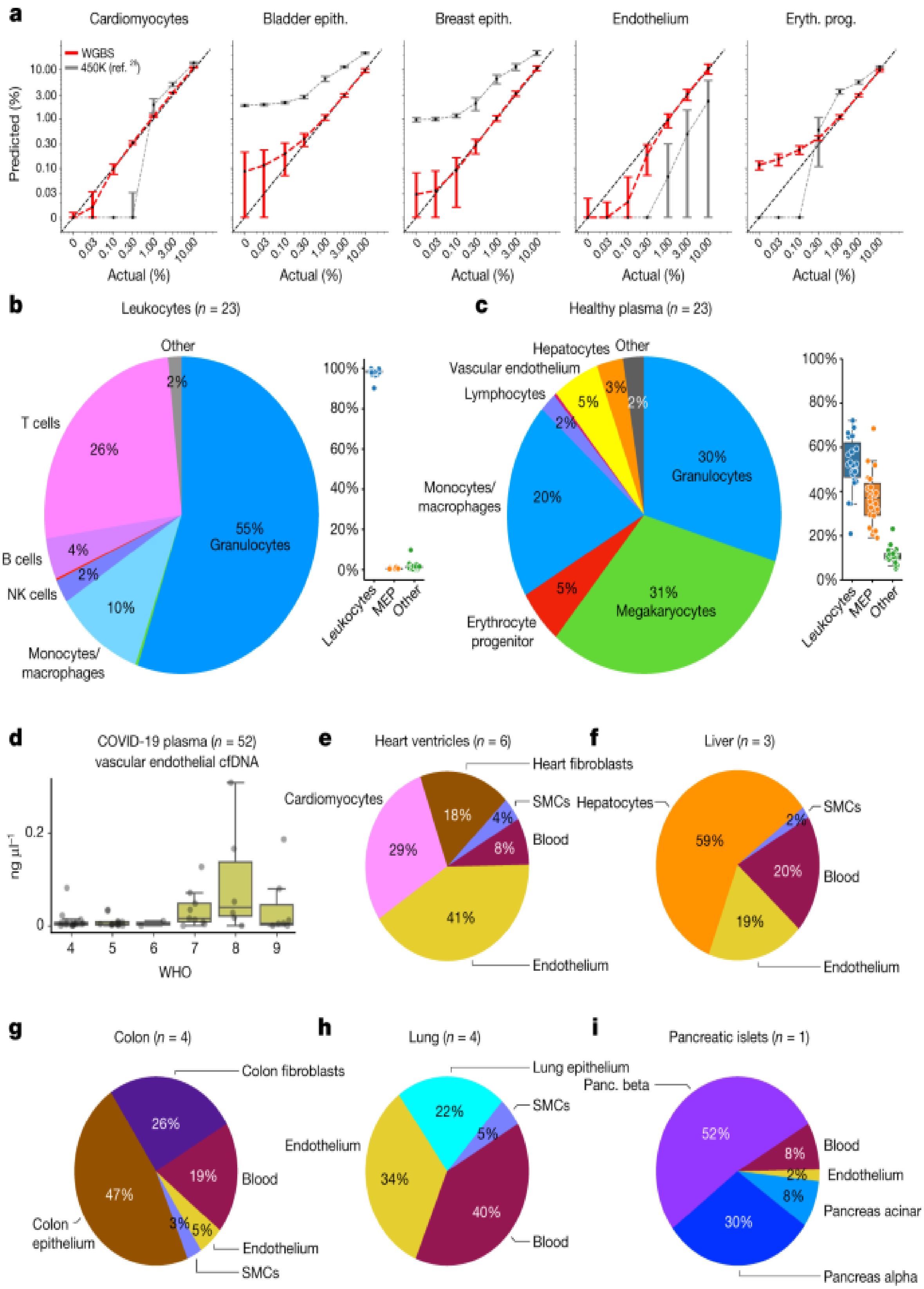
Fragment-level methylome deconvolution

Last, we developed a computational fragment-level deconvolution algorithm for DNA methylation sequencing data and used the top 25 markers defined

for each cell type (a total of 1,246 markers) to study methylomes obtained from composite tissue samples and cfDNA. Briefly, we generated an atlas in which the percentage of unmethylated fragments is computed for every marker (row) in each cell type (column). A non-negative least-squares (NNLS) algorithm is then used to fit an input sample and estimate its relative contributions (Supplementary Information).

To estimate the accuracy of our fragment-level approach, we used in silico mixtures of sequenced reads. For each cell type we applied a leave-one-out approach to mix one held-out sample in leukocyte reads, then used the deconvolution algorithm to infer cellular composition in the mixture. We repeated this process at concentrations varying from 0 to 10%. As shown in Fig. 6a, we found that the 1,246 markers (top 25 per cell type) allowed accurate detection of DNA from a given source at around 0.1% resolution, an improvement of nearly one order of magnitude in comparison with array-based approaches[28]. Four-way in silico mixes, in which endothelial and hepatocyte methylomes were also included to realistically mimic cfDNA composition, yielded similar results (Extended Data Fig. 8).

Fig. 6: Fragment-level deconvolution using cell type-specific biomarkers.


a
Cardiomyocytes
Bladder epith.
Breast epith.
Endothelium
Eryth. prog.
WGBS
450K (ref. 26)
Predicted (%)
Actual (%)
b
Leukocytes (n = 23)
Other
T cells
B cells
NK cells
Monocytes/ macrophages
Granulocytes
55%
26%
10%
4%
2%
2%
Leukocytes
MEP
Other
c
Healthy plasma (n = 23)
Hepatocytes
Vascular endothelium
Lymphocytes
Monocytes/ macrophages
Erythrocyte progenitor
Granulocytes
Megakaryocytes
30%
31%
20%
5%
5%
3%
2%
2%
d
COVID-19 plasma (n = 52) vascular endothelial cfDNA
ng µl−1
WHO
e
Heart ventricles (n = 6)
Heart fibroblasts
Cardiomyocytes
SMCs
Blood
Endothelium
18%
29%
4%
8%
41%
f
Liver (n = 3)
Hepatocytes
SMCs
Blood
Endothelium
59%
2%
20%
19%
g
Colon (n = 4)
Colon fibroblasts
Blood
Colon epithelium
Endothelium
SMCs
26%
19%
47%
3%
5%
h
Lung (n = 4)
Lung epithelium
SMCs
Endothelium
Blood
22%
5%
34%
40%
i
Pancreatic islets (n = 1)
Panc. beta
Blood
Endothelium
Pancreas acinar
Pancreas alpha
52%
8%
2%
8%
30%

a, Cell-type-specific markers achieved less than 0.1% resolution. In silico simulations for five cell types, in which held-out samples were computationally mixed within leukocytes then analysed using 1,246 atlas markers plus 25 additional megakaryocyte markers (red) or an array-based deconvolution of these mixes[28] (grey). Box plots show average contribution in ten simulations, with error bars representing 1 s.d. **b**,**c**, Cell type composition in leukocytes (**b**) and plasma samples (**c**) from healthy donors. Box plots show overall proportions of leukocytes, megakaryocytes and erythroblasts (MEP) and other cell types. **d**, Analysis of low-coverage plasma samples from 52 patients with SARS-CoV-2 (ref. [44]) identified endothelial-derived cfDNA in patients with WHO ordinal scale seven or higher (requiring admittance to intensive care unit). **e**–**i**, Fragment-level deconvolution of Roadmap/ENCODE samples[4,5] showing cell-type-specific contributions. **e**, Heart ventricle samples contained a mixture of cardiomyocytes, endothelial cells, fibroblasts and blood. **f**, Liver samples contained around 60% of hepatocyte DNA, plus blood and endothelial cells. **g**, Colon samples contained approximately 50% epithelium, plus fibroblasts and blood. **h**, Lung samples contained less than 30% of lung epithelial cells. **i**, Pancreatic islet samples contained beta, alpha, duct and acinar cells. Box plots denote median and IQR, with whiskers 1.5× IQR.

We then estimated the cellular composition of leukocytes and cfDNA using WGBS data from 23 healthy donors; 99.5% of leukocyte-derived DNA was attributed to granulocytes, monocytes, macrophages and NK, T and B cells, consistent with typical blood counts (Fig. 6b and Supplementary Table 8). The cfDNA of healthy subjects was mostly derived from leucocytes: granulocytes (29.7%), monocytes/macrophages (20%) and lymphocytes (3%). Solid tissues contributing to cfDNA included vascular endothelial cells (6%) and hepatocytes (3.1%) (Fig. 6c), consistent with previous results[28]. The current atlas also shows a significant contribution of megakaryocytes (31%) and erythrocyte progenitor (prog.) cells (5%) to cfDNA, which were not observed in previous studies that used reference methylomes of a more limited scope.

Endothelial cfDNA in patients with COVID-19

Analysis based on DNA methylation patterns offers an opportunity to identify the tissue origins of cfDNA. COVID-19 inflicts damage to multiple tissues, some of which have no biomarkers. We used the atlas to deconvolve shallow WGBS data from 52 patients hospitalized owing to COVID-19 (ref. [44]). We identified excessive cell-free DNA fragments from granulocytes, erythrocyte progenitors, lung and liver, consistent with published analysis of these samples (Supplementary Information). Strikingly, we also identified a significant contribution of vascular endothelial cells to the cfDNA of these patients, which could not be detected in the published analysis in the absence of an endothelial cell methylome reference (Fig. 6d). Interestingly, the concentration of endothelial cell-derived cfDNA was higher in patients with severe disease (WHO score ≥7) compared with those with milder disease (WHO score ≤6; $P \leq 6 \times 10^{-5}$, Mann–Whitney). These results suggest that vascular endothelial cell death plays a substantial role in the pathogenesis of COVID-19, potentially related to coagulopathy, and highlight the benefit of using a comprehensive cell-type-specific atlas for cfDNA methylome analysis.

Cell type deconvolution of composite tissues

Finally, we analysed whole-genome methylomes from ENCODE[5] and the Roadmap Epigenomics atlas[4] using our atlas (based on 25 markers per cell type). Deconvolution of some methylomes showed a homogenous composition as intended—for example, 97–99% T cell DNA in Roadmap T cell samples (Supplementary Table 9). However, analysis of other samples showed a highly heterogeneous composition, as previously reported based on array-based bulk tissue deconvolution algorithms such as EpiDISH and EpiScore[14,15,45]. For example, heart ventricle samples comprised 29% cardiomyocytes, 41% endothelial cells and 18% cardiac fibroblasts (Fig. 6e); liver methylomes comprised around 60% hepatocytes, 21% blood and 20% endothelial cells; and colon methylomes comprised about 50% colon epithelium, 26% colon fibroblasts and 19% blood. Most strikingly, Roadmap lung samples were dominated by blood (40%), endothelium (34%) and smooth muscle (5%), with only 22% of DNA derived from lung epithelial cells (Fig. 6f–i and Supplementary Table 9). Importantly, a similar deconvolution of the 205 samples presented here yielded an average

contribution of 94% for the expected cell type for each sample (median of 95%, Supplementary Table 10), or of 91% (median of 92%) in a more stringent leave-one-out cross-validation analysis (Supplementary Table 11), highlighting the purity of collected samples.

Naturally, fragment-level analysis is limited to cell types for which whole-genome sequencing data are available, and some cell types can be analysed only by array-based algorithms[15,28]. Nonetheless, the markers and algorithm presented here allow analysis of composite bulk tissue and plasma samples, across multiple cell types and with high accuracy.

Discussion

The comprehensive atlas of human cell type methylomes described here sheds light on principles of DNA methylation and provides a valuable resource for multiple lines of investigation, as well as translational applications.

Our analysis used whole-genome sequencing data to show that methylation patterns are strikingly similar among healthy replicates of the same cell type from different individuals. The similarity between individuals reflects the robustness of cell differentiation and maintenance circuits, at least as far as healthy tissues are concerned. Pathologies involving destabilization of the epigenome obviously disrupt these circuits, resulting in a larger variety of methylation patterns among cells descended from a specific normal cell type. We predict that, even in cancers (of the same primary anatomic site and histologic type), comparative methylome analysis of purified epithelial cells, performed at the level of methylation blocks, will show a smaller interindividual variation than typically assumed.

As the atlas demonstrates, each cell type has a set of genomic regions that are uniquely unmethylated in that cell type compared with others, as well as additional genomic regions that share methylation patterns with related cell types. Using unsupervised clustering of cell-type-specific methylomes, we found that cell types were clustered in ways that reflected their developmental origins rather than expression patterns. This offers a fascinating view of DNA methylation as a record of the methylomes of

progenitor cells, retained in the genome through dramatic developmental transitions and decades of life thereafter. We propose that comparative methylome analysis will allow reconstruction of parts of the methylomes of fetal structures or cell types, similarly to the reconstruction of last common ancestors in evolutionary biology.

The vast majority of cell-type-specific differentially methylated regions were specifically demethylated in one cell type. The chromatin of these regions is typically highly accessible and bears histone marks associated with active gene regulation, as found in enhancers and promoters. Moreover, these loci are enriched for TF binding site motifs that operate in that cell type. We devised an integrated approach that, based on distance and gene expression profiles, allowed us to highlight potential target genes for these putative enhancer regions. Many enhancer regions were associated with nearby genes that are broadly expressed, potentially reflecting gene regulation by multiple tissue-specific enhancers. Our findings are consistent with previous studies that showed tissue-specific hypomethylation occurring at gene enhancers[35,36,37]. Our data-driven approach for marker identification is complementary to recent gene-centric approaches[14,15] that use tissue-specific single-cell RNA sequencing data to define marker genes and identify neighbouring CpGs specifically unmethylated in target cell types. Finally, we devised a fragment-level genomic analysis to identify tens of thousands of unmethylated regions, per cell type, which were annotated with genomic features, DNA accessibility, chromatin marks and TF binding motifs to produce a cell type-specific catalogue of putative enhancers. Further analysis of this atlas will show and validate the complete set of human enhancers in each cell type.

Conversely, we identified genomic regions specifically methylated in one or two cell types, representing around 3% of cell-type-specific differentially methylated regions. These are often located in CpG islands and characterized by H3K27me3 and Polycomb binding in tissues where the locus is not methylated[40,41]. This epigenetic repressive switching was previously described in cancer and during early development[41,46], but its role during differentiation of specific cell types remains unclear. These regions are enriched for CTCF binding sites, suggesting a role for DNA methylation in attenuating the binding of CTCF and thus modulation of the

cell-type-specific, three-dimensional organization of neighbouring DNA[35,36,47].

For DNA methylation sequencing data, the atlas described here is, to our knowledge, the most comprehensive compendium to date. We identified more than one thousand cell-type-unique DNA methylation regions that could serve as accurate and specific biomarkers for fragment-level analysis and identification of cell death events by monitoring of cfDNA. Notably, most of these marker regions are not covered by 450K/EPIC BeadChip DNA methylation arrays, and were not previously appreciated. To allow interpretation of array data, we offer alternative sets of cell-type-specific markers limited to CpG sites included in BeadChip 450K arrays. Similarly, we identified cell-type-specific markers in regions targeted by both RRBS and hybrid capture panels (Extended Data Fig. 9 and Supplementary Tables 12–17). As shown in Extended Data Fig. 10, the array-adapted atlas allows high-resolution interpretation of array methylomes of pancreatic islet, lung and breast biopsies, highlighting the presence of cell types not previously profiled[48,49,50].

Many cell types are missing from the atlas, typically because of limited availability of material. Examples include osteoblasts, cholangiocytes, cells of the adrenal gland, urethral epithelium and haematopoietic stem cells. Additionally, we did not separate many subpopulations of interest—for example, different types of neurons or lymphocytes. The atlas is viewed as a living, publicly available database to be updated in the future. The resolution of the atlas yields a quantitative understanding of composite tissues and allows one to identify missing methylomes of additional cell types yet to be characterized. We also acknowledge that the purity of the sorted cell populations varies, owing to variation in the quality of antibodies used for FACS and the extent to which they allow separation of cell types. Nonetheless, even the least pure cell types in the atlas (for example, some preparations of vascular endothelial cells, fibroblasts, SMC and adipocytes showing 70–80% purity), when averaged over replicates, are useful for identification of differentially methylated regions and for inference of cell composition in mixtures.

In summary, we present a comprehensive methylation atlas of primary human cell types along with an extensive set of cell-type-specific markers

and computation tools for fragment-level analysis of mixed cell type samples. These complement the plethora of array-based methylomes and deconvolution tools available for the analysis of array data. Together, the data shed light on the roles of DNA methylation in cellular biology and gene regulation and facilitate the identification of enhancers active in each cell type. Perhaps the most promising utility of our atlas is the potential for fragment-level deconvolution of mixed cell type samples, allowing sensitive identification of the tissue of origin of cfDNA in plasma of individuals with cancer and other diseases[18,19,20,21,28,29,30].

Methods

Human tissue samples

Human tissues were obtained from various sources, as detailed in Supplementary Table 1. The majority (148) of the 205 samples analysed were sorted from tissue remnants obtained at the time of routine, clinically indicated surgical procedures at the Hadassah Medical Center. In all cases, normal tissue distant from any known pathology was used. Surgeons and/or pathologists were consulted before removal of tissue to confirm that its removal would not compromise the final pathologic diagnosis in any way. For example, in patients undergoing right colectomy for carcinoma of the caecum, the most distal part of the ascending colon and most proximal part of the terminal ileum were obtained for cell isolation. Normal bone marrow was obtained at the time of joint replacement in patients with no known haematologic pathology. The patient population included 135 individuals (n = 60 males, n = 74 females) aged 3–83 years. The majority of donors were White. Approval for collection of normal tissue remnants was provided by the Institutional Review Board (IRB, Helsinki Committee), Hadassah Medical Center, Jerusalem, Israel. Written informed consent was obtained from each donor or legal guardian before surgery.

As described in Supplementary Table 1, some cells and tissues were obtained through collaborative arrangements: pancreatic exocrine and liver samples (cadaveric organ donors, n = 5) from M. Grompe, Oregon Health & Science University; adipocytes (subcutaneous adipocytes at time of cosmetic surgery following weight loss, n = 3), oligodendrocytes and neurons (brain

autopsies, $n = 14$) from K. L. Spalding and H. Druid, Karolinska Institute, Stockholm; and research-grade cadaveric pancreatic islets from J. Shapiro, University of Alberta ($n = 16$). In all cases, tissues were obtained and transferred in compliance with local laws and after the approval of the local ethics committee on human experimentation. Sixteen cell types were obtained from commercial sources, including 15 from Lonza and one from Sigma-Aldrich. Three pancreatic islet preparations were obtained from the Integrated Islet Distribution Program (https://iidp.coh.org).

Tissue dissociation and FACS sorting of purified cell populations

Fresh tissue obtained at the time of surgery was trimmed to remove extraneous tissue. Cells were dispersed using enzyme-based protocols optimized for each tissue type. The resulting single-cell suspension was incubated with the relevant antibodies and FACS sorted to obtain the desired cell type (Extended Data Fig. 2 and Supplementary Information).

Purity of live sorted cells was determined by messenger RNA analysis for key known cell-type-specific genes, whereas the purity of cells fixed before sorting was determined using previously validated cell-type-specific methylation signals (Extended Data Fig. 2c and Supplementary Information). DNA was extracted using the DNeasy Blood and Tissue kit (no. 69504, Qiagen) according to the manufacturer’s instructions, and stored at −20 °C for bisulfite conversion and whole-genome sequencing.

WGBS

Up to 75 ng of sheared genomic DNA was subjected to bisulfite conversion using the EZ-96 DNA Methylation Kit (Zymo Research), with liquid handling on a MicroLab STAR (Hamilton). Dual-indexed sequencing libraries were prepared using Accel-NGS Methyl-Seq DNA library preparation kits (Swift BioSciences) and custom liquid handling scripts executed on the Hamilton MicroLab STAR. Libraries were quantified using KAPA Library Quantification Kits for Illumina Platforms (Kapa Biosystems). Four uniquely dual-indexed libraries, along with the 10% PhiX v.3 library (Illumina), were pooled and clustered on an Illumina NovaSeq

6000 S2 flow cell followed by 150 bp, paired-end sequencing. Total read count and average sequencing depth (in read pairs), as well as percentage of CpGs, per sample, at 1× and 10×, are detailed in Supplementary Table 1. Also listed are average methylation levels, per sample, at CpG, nonCpG and CC dinucleotides. Intriguingly, sorted neuron samples showed higher CpA methylation (approximately 10%) compared with other samples (approximately 1%).

WGBS computational processing

Paired-end FASTQ files were mapped to the human (hg19, hg38), lambda, pUC19 and viral genomes using bwa-meth (v.0.2.0)[51] then converted to BAM files using SAMtools (v.1.9)[52]. Duplicated reads were marked by Sambamba (v.0.6.5) with parameters '-l 1 -t 16 --sort-buffer-size 16000 --overflow-list-size 10000000' (ref. [53]). Reads with low mapping quality, duplicated or not mapped in a proper pair were excluded using SAMtools view with parameters '-F 1796 -q 10'. Reads were stripped from nonCpG nucleotides and converted to PAT files using wgbstools (v.0.1.0)[54].

Genomic segmentation into multisample homogenous blocks

We developed and implemented a multichannel dynamic Pprogramming segmentation algorithm to divide the genome into continuous genomic regions (blocks), showing homogeneous methylation levels across multiple CpGs for each sample[54]. A generative probabilistic model is used, each block inducing a Bernoulli distribution with some ${\theta }_{i}^{k}$, where i is the block index and k the sample index ($k = 1,..., K$), and each observation (occurence of one CpG at one sequenced fragment) is represented by a random variable sampled i.i.d. (independent and identically distributed) from the same beta value Ber ${\theta }_{i}^{k}$. The log-likelihood of all sequencing data is the sum of log-likelihoods across all blocks, each decomposing as the sum of log-likelihoods across all samples. The log-likelihood of the ith block can therefore be formalized as:

$${\rm{score}}({{\rm{block}}}_{i})={ll}_{i}={{\varSigma }^{K}}_{k=1}({{({N}_{C})}_{i}}^{k}\times \log ({{\hat{\theta

}}_{i}}^{k})+{{({N}_{T})}_{i}}^{k}\times \log (1-{{\hat{\theta }}_{i}}^{k}))$$

where \({({N}_{C})}_{i}^{k}\,,\,{({N}_{T})}_{i}^{k}\) is the number of methylated and unmethylated observations, respectively, in the *i*th block in the *k*th sample, whereas \({{\hat{\theta }}_{i}}^{k}\) marks a Bayes estimator of the Bernoulli distribution parameter, calculated with \({a}_{C},{a}_{T}\) pseudocounts for each block/sample:

$${{\hat{\theta }}_{i}}^{k}=\frac{{{({N}_{C})}_{i}}^{k}+{\alpha }_{C}}{{{({N}_{C})}_{i}}^{k}+{{({N}_{T})}_{i}}^{k}+{\alpha }_{C}+{\alpha }_{T}}$$

These hyperparameters are used for regularization, to control the trade-off between overfitting (shorter blocks) and generalization (longer blocks). Dynamic programming is then used to find the optimal segmentation across the genome. Briefly, we maintain a 1 × *N* table *T* (*N* = 28,217,448 CpGs) for optimal segmentation scores across all prefixes. Specifically, *T*[*i*] holds the score of the optimal segmentation of all CpG sites from 1 through to i, and *T*[*N*] holds the final, optimal, score across the entire genome. The table itself is updated sequentially from 1 to *N*, where the optimal segmentation up to position *i* is achieved by the addition of a new block to a shorter optimal segmentation (for example, up to position *i*′):

$$T[i]=\mathop{\max }\limits_{i{\prime} < i}\{T[i{\prime}]+{\rm{score}}({\rm{block}}[i{\prime} +1,...,i])\}$$

For this, all previous optimal segmentations are considered and a new block is added from position ($i' + 1$) to position *i* (with a maximal block size of 5,000 bp). The combination that maximizes log-likelihood is selected as the optimal segmentation from 1 to *i*, and the start index of the last block is recorded in a traceback table. Once the score of optimal segmentation is calculated in *T*[*N*], the traceback table is used to retrieve the full segmentation. An upper bound on block length (5,000 bases) is set to improve running times and each chromosome is run separately. The linear distance between consecutive CpGs is ignored under this model. The model and segmentation algorithm are further described in Supplementary Information.

Segmentation and clustering analysis

We segmented the genome into 7,104,162 blocks using wgbstools (with parameters 'segment --max_bp 5000') with all of the 205 samples as reference, and retained 2,099,681 blocks covering at least four CpGs. For hierarchical clustering (Fig. 2) we selected the top 1% (20,997) blocks showing the highest variability in average methylation across all samples. Blocks with sufficient coverage of at least ten observations (calculated as sequenced CpG sites) across two-thirds of the samples were further retained. We then computed the average methylation for each block and sample calculated using wgbstools (--beta_to_table -c 10), marked blocks with fewer than ten observations as missing values and imputed their methylation values using sklearn KNNImputer (v.0.24.2)[55]. The 205 samples were clustered with the unsupervised agglomerative clustering algorithm[23], using scipy (v.1.6.3)[56] and L1 norm. The fanning diagram was plotted using ggtree (v.2.2.4)[57].

Cell-type-specific markers

The 205 atlas samples were divided into 51 groups by cell type, yielding 39 basic groups and 12 composite supergroups (Supplementary Table 3). We then performed a one-versus-all comparison to identify differentially methylated blocks unique for each cell type. For this we used wgbstools' 'find_markers' function to first identify blocks covering at least five CpGs (length 10–1,500 bp) to calculate the average methylation per block/sample and rank the blocks according to the difference in average methylation between target samples versus all other samples. To allow some flexibility, this difference was computed (for unmethylated markers) as the difference between the 75th percentile in target samples (typically allowing one outlier) versus the 2.5th percentile in the background group (typically allowing about five outlier samples). For methylated markers, this was computed as the difference between the 25th and 97.5th percentiles (Supplementary Information). Low-coverage blocks (fewer than 25 observations), in which the estimation error of average methylation was around 10%, were replaced by a default value of 0.5 which is neither unmethylated nor methylated, thus reducing the block's methylation difference and downgrading its rank. For

cell type-specific markers we selected the top 25 per cell type, for a total of 1,246 markers (Supplementary Table 4a).

Atlases for 450K/EPIC, RRBS and hybrid capture panels were identified similarly while examining a subset of genomic regions, overlapping various probe sets or genomic regions (-b option). Chromatin analysis was performed on the top 250 markers per cell type (total of 11,713 markers; Supplementary Table 4b). Motif analysis was performed on the top 1,000 markers per cell type (total of 50,286 markers; Supplementary Table 4b) using the difference between the 25th and 75th percentile, to allow putative enhancers unmethylated in additional cell types.

Enrichment for gene set annotations

Analysis of gene set enrichment was performed using GREAT[31]. For each cell type we selected the top 250 differentially unmethylated regions and ran GREAT via batch web interface using default parameters. Enrichments for 'Ensembl Genes' were ignored, and a significance threshold of binomial false discovery rate ≤0.05 was used.

Enrichment for chromatin marks

For each cell type we analysed the top 250 differentially unmethylated regions versus published ChIP–seq (H3K27ac and H3K4me1) and DNase sequencing from the Roadmap Epigenomics project (downloaded from ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/by_experiment and http://egg2.wustl.edu/roadmap/data/byDataType/dnase/BED_files_enh) in bigWig and bed formats. These include E032 for B cell markers, E034 for T cell markers, E029 for monocyte/macrophage markers, E066 for liver hepatocytes, E104 for heart cardiomyocytes and fibroblasts and E109 and E110 for gastric/small intestine/colon[4]. Annotations for chromHMM were downloaded (15-states version) from https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final[3], and genomic regions annotated as enhancers (7_Enh) were extracted and reformatted in bigWig format. Raw single-cell ATAC–seq data were downloaded from GEO GSE165659 (ref. [32]) as 'feature' and 'matrix' files for 70 samples. For each sample, cells of

the same type were pooled to output a bedGraph file, which was mapped from hg38 to hg19 using UCSC liftOver[58]. Overlapping regions were dropped using bedtools (v.2.26.0)[59]. Finally, bigWig files were created using bedGraphToBigWig (v.4)[60]. Heatmaps and average plots were prepared using deepTools (v.3.4.1)[61], with the functions 'computeMatrix', 'plotHeatmap' and 'plotProfile'. We used default parameters except for 'referencePoint=center', 15 kb margins and 'binSize=200' for ChIP–seq, DNaseI and chromHMM data, and 75 kb margins with 'binSize=1000' for ATAC–seq data.

Motif analysis

For each cell type we analysed the top 1,000 differentially unmethylated regions for known motifs (Supplementary Table 6a) using the HOMER function 'findMotifsGenome.pl', with parameters '-bits' and '-size 250'[39]. Similar analyses were performed for the unmethylated regions in each cell type (Supplementary Table 6b), as well as unmethylated regions overlapping H3K27ac, but not H3K4me3, peaks (Supplementary Table 6c).

Methylation marker–gene associations

For each cell-type-specific marker we identified all neighbouring genes up to 500 kb apart. We then examined the expression levels of these genes across the GTEx dataset covering 50 tissues and cell types[62]. We then standardized the expression of each gene across all conditions, by replacing expression values with standard deviations (*z*-scores) above/below the average expression of that gene across samples. This was followed by column-wise standardization in which the relative enrichment of a gene under a given condition is normalized by the enrichment of other genes under that condition. This highlighted the most overexpressed genes for each tissue. We then classified each 'marker–gene–condition' combination as tier 1: distance ≤5 kb, expression ≥10 TPM and *z*-score ≥1.5; tier 2: same as tier 1 but with distance ≤50 kb; tier 3: up to 750 kb, expression ≥25 TPM and *z*-score ≥5; and tier 4: same as tier 3 but with *z*-score ≥3.5.

A catalogue of unmethylated loci and putative enhancers for each cell type

For each genomic region (blocks of at least four CpGs), and for any of the 39 cell type groups, fragments with at least four CpGs from all replicates were merged and classified as either U (fragment-level methylation 15% or less), M (at least 85%) or X (over 15% but below 85%). The percentage of U fragments was then calculated using ‘wgbstools homog --threshold .15,.85’, and blocks with at least 85% unmethylated fragments retained. These blocks were overlapped with genomic features based on UCSC hg19 annotations, including CpG islands and transcriptional start site regions (up to 1 kb from a gene start site). We also used narrowPeak annotations downloaded from Roadmap[4] and ENCODE project[5] (accessions listed in Supplementary Table 6d). hg38 bed files were converted to hg19 using liftOver[58]. For putative enhancers, nonpromoter active regulatory regions were defined as those overlapping H3K27ac, but not H3K4me3, peaks under matching conditions. TF binding sites were downloaded from JASPAR 2022 (ref. [63]).

Interindividual variation in cell type methylation

We define a similarity score between two samples as the fraction of blocks containing at least three CpGs and at least ten binary observations (sequenced CpG sites) in which the average methylation of the two samples differs by at least 0.5. Only cell types with $n \geq 3$ FACS-sorted replicates from different donors are considered (136 samples in total).

CTCF ChIP–seq analysis

CTCF ChIP–seq data were downloaded from the ENCODE project[5] as 168 bigWig files, covering 61 tissues/cell types (hg19). Samples of the same cell type were averaged using multiBigwigSummary (v.3.4.1)[61].

Endodermal marker analysis

All 892 endodermal hypomethylated markers were found using wgbstools function ‘find_markers’ (v.0.2.0), with parameters ‘--delta_quants 0.4 --tg_quant 0.1 --bg_quant 0.1’ (ref. [54]). For endoderm-derived epithelium, 51 samples were compared with 103 nonepithelial samples from mesoderm or ectoderm. Blocks were selected as markers if the average methylation of the 90th percentile of the epithelial samples was lower than the tenth percentile of the nonepithelial samples by at least 0.4.

UXM fragment-level deconvolution algorithm

We developed a fragment-level deconvolution algorithm: each fragment was annotated as U (mostly unmethylated), M (mostly methylated) or X (mixed) depending on the number of methylated and unmethylated CpGs[64]. We then calculated, for each genomic region (marker) and across all cell types, the proportion of U/X/M fragments with at least k CpGs. Here we used $k = 4$ and thresholds of less than or equal to 25% methylated CpGs for U reads, and more than or equal to 75% methylated CpGs for M reads. We then constructed reference atlas A with 1,232 regions (top 25 markers per cell type), in which the $A_{i,j}$ cell holds the U proportion of the ith marker in the jth cell type. Given an input sample, the U proportion at each marker is computed to form a 1,232 × 1 vector b. Then, NNLS is applied to infer coefficient vector x by minimizing ${| A\times x-b| }_{2}$ subject to non-negative x, normalized to ${\Sigma }_{j}{x}_{j}=1$. Alternatively, each marker can be weighed differently based on fragment coverage in the input sample. For this, b can be defined as the number of U fragments in each region and the rows of A similarly multiplied by Ci, the total number of fragments in each region, thus minimizing ${| {\rm{diag}}(C)\times A\times x-b| }_{2}$. Additional details are available in Supplementary Information.

In silico simulation of WGBS deconvolution

Simulated mixtures were performed for cardiomyocytes ($n = 4$), bladder epithelium ($n = 5$), breast epithelium ($n = 7$), endothelial cells ($n = 19$) and erythrocyte progenitors ($n = 3$) in a leave-one-out manner. For this, one sample was held out and segmentation and marker selection (25 per cell type) were rerun using the remaining 204 samples. We then simulated

mixtures by sampling and mixing reads from the held-out sample at 10, 3, 1, 0.3, 0.1, 0.03 and 0% into a background of leukocyte samples. This was repeated ten times. Finally, mixed samples were analysed using the UXM fragment-level algorithm with markers from the reduced (204) atlas, using fragments with at least three CpGs. Merging, splitting and mixing of reads were performed using wgbstools (v.0.1.0)[54].

Array-based analysis was performed by computing, for each mixed set of fragments, average methylation levels across each of around 480,000 CpG sites present in the 450K array ('wgbstools beta_to_450k'). We then deconvolved these data according to the method of Moss et al.[28] (https://github.com/nloyfer/meth_atlas).

We also simulated four-way mixtures in which background plasma methylomes were simulated as a combination of 90% fragments from leukocytes, 7.5% from a vascular endothelial sample and 2.5% from a hepatocyte sample. As described above, this was done by holding out the three samples (for example, cardiomyocytes, endothelial cells and hepatocytes) and then rererunning segmentation and marker selection on the (202 = 205 – 3) remaining samples, to obtain a set of markers that was then used for fragment-level deconvolution of mixtures.

WGBS deconvolution

Leukocytes and matching plasma samples ($n = 23$) were processed as described above and analysed using the WGBS methylation atlas, including 1,246 markers plus (for plasma samples) an additional 25 megakaryocyte markers. Fifty-two plasma samples from 28 patients with SARS-CoV-2 (ref. [44]) downloaded as FASTQ files were processed as described above. Because of the low coverage (1–2×) of these samples, we extended the marker set from the top 25 to the top 250 markers per cell type (Supplementary Table 4b), and also included 250 megakaryocyte markers[65]. Roadmap[4] and ENCODE[5] samples were processed as described above and analysed using the UXM algorithm.

Deconvolution of 450K array data

Previously published 450K array data were downloaded from either The Cancer Genome Atlas (lung and breast biopsies)[49,50] or GEO accession no. GSE62640 (ref. [48]) and deconvoluted with meth_atlas NNLS software (https://github.com/nloyfer/meth_atlas) using our array-adapted atlas (Supplementary Table 12). Breast biopsies were grouped using PAM50 classifications[66].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

DNA methylation data are available in formats bigWig (position and average methylation across 28,217,448 CpGs) and beta (a similar wgbstools-compatible binary format) at GEO (accession no. GSE186458). BigWig and beta files for hg38 are also available. Fragment-level information (in pat format, including CpG starting index, methylation pattern of all covered CpGs and number of fragments with exact multiCpG pattern) are also available. Raw fastq files have been deposited at the European Genome-phenome Archive (EGA) under study accession number: EGAS00001006791 and can be downloaded upon request to EGA (through the atlas Data Access Committee).

Code availability

Code is available at github.com/nloyfer/wgbs_tools and github.com/nloyfer/UXM_deconv.

References

1. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine. *Lancet* **392**, 777–786 (2018).

2. Ziller, M. J. et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).

3. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

4. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

5. ENCODE Project Consortiumet al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

6. Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE* **7**, e41361 (2012).

7. Smith, Z. D. et al. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).

8. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).

9. Titus, A. J., Gallimore, R. M., Salas, L. A. & Christensen, B. C. Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum. Mol. Genet.* **26**, R216–R224 (2017).

10. Bibikova, M. et al. High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).

11. Lehmann-Werman, R. et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl Acad. Sci. USA* **113**, E1826–E1834 (2016).

12. Guo, S. et al. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **49**, 635–642 (2017).

13. Franzen, J. et al. DNA methylation changes during long-term in vitro cell culture are caused by epigenetic drift. *Commun. Biol.* **4**, 598 (2021).

14. Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol.* **21**, 221 (2020).

15. Zhu, T. et al. A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. *Nat. Methods* **19**, 296–306 (2022).

16. Wan, J. C. M. et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017).

17. Aravanis, A. M., Lee, M. & Klausner, R. D. Next-generation sequencing of circulating tumor DNA for early cancer detection. *Cell* **168**, 571–574 (2017).

18. Sun, K. et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl Acad. Sci. USA* **112**, E5503–E5512 (2015).

19. Kang, S. et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol.* **18**, 53 (2017).

20. Li, W. et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res.* **46**, e89 (2018).

21. Klein, E. A. et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann. Oncol.* **32**, 1167–1177 (2021).

22. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

23. Michener, C. D. & Sokal, R. R. A quantitative approach to a problem in classification. *Evolution* **11**, 130–162 (1957).

24. Gu, G., Dubauskaite, J. & Melton, D. A. Direct evidence for the pancreatic lineage: NGN3$^+$ cells are islet progenitors and are distinct from duct progenitors. *Development* **129**, 2447–2457 (2002).

25. Le Douarin, N. M. On the origin of pancreatic endocrine cells. *Cell* **53**, 169–171 (1988).

26. van Arensbergen, J. et al. Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome Res.* **20**, 722–732 (2010).

27. Frank, D. B. et al. Early lineage specification defines alveolar epithelial ontogeny in the murine lung. *Proc. Natl Acad. Sci. USA* **116**, 4362–4371 (2019).

28. Moss, J. et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).

29. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).

30. Gai, W. et al. Applications of genetic-epigenetic tissue mapping for plasma DNA in prenatal testing, transplantation and oncology. *eLife* **10**, e64356 (2021).

31. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

32. Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001 (2021).

33. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

34. Kirillov, A. et al. A role for nuclear NF-kappaB in B-cell-specific demethylation of the Igkappa locus. *Nat. Genet.* **13**, 435–441 (1996).

35. Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).

36. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).

37. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).

38. Feldmann, A. et al. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.* **9**, e1003994 (2013).

39. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

40. Straussman, R. et al. Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.* **16**, 564–571 (2009).

41. Gal-Yam, E. N. et al. Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc. Natl Acad. Sci. USA* **105**, 12979–12984 (2008).

42. Wang, H. et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).

43. Wiehle, L. et al. DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.* **29**, 750–761 (2019).

44. Cheng, A. P. et al. Cell-free DNA tissues of origin by methylation profiling reveals significant cell, tissue, and organ-specific injury related to COVID-19 severity. *Med (N Y)* **2**, 411–422 (2021).

45. Zheng, S. C. et al. A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics* **10**, 925–940 (2018).

46. Li, Y. et al. Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol.* **19**, 18 (2018).

47. Nordin, M., Bergman, D., Halje, M., Engström, W. & Ward, A. Epigenetic regulation of the Igf2/H19 gene cluster. *Cell Prolif.* **47**, 189–199 (2014).

48. Hall, E. et al. Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. *Genome Biol.* **15**, 522 (2014).

49. Weisenberger, D. J. Characterizing DNA methylation alterations from The Cancer Genome Atlas. *J. Clin. Invest.* **124**, 17–23 (2014).

50. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

51. Pedersen, B. S., Eyring, K., De, S., Yang, I. V. & Schwartz, D. A. Fast and accurate alignment of long bisulfite-seq reads. Preprint at https://doi.org/10.48550/arXiv.1401.1129 (2014).

52. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

53. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).

54. Loyfer, N. & Kaplan, T. wgbstools: A computational suite for DNA methylation sequencing data representation, visualization, and analysis. https://github.com/nloyfer/wgbs_tools (accessed 28 November 2022).

55. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

56. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

57. Yu, G. Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).

58. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

59. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

60. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).

61. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

62. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

63. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).

64. Loyfer, N. & Kaplan, T. UXM: a fragment-level algorithm for computational deconvolution of whole tissues and plasma samples using DNA methylation sequencing data. https://github.com/nloyfer/UXM_deconv (accessed 28 November 2022).

65. Moss, J. et al. Megakaryocyte and erythroblast DNA in plasma and platelets. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.03.510502 (2022).

66. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).

Acknowledgements

We thank H. Cedar and N. Friedman for insightful discussions. We also thank members of the Dor, Kaplan and Rosenfeld laboratories. This work was supported by grants from GRAIL, Alzheimer's Drug Discovery Foundation, Human Islet Research Network (nos. HIRN UC4DK116274 and UC4DK104216), the Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine, The Alex U Soyka pancreatic cancer fund, The Israel Science Foundation, the Waldholtz/Pakula family, the Robert M. and Marilyn Sternberg Family Charitable Foundation, the Helmsley Charitable Trust and DON Foundation (to Y.D.), Israel Science Foundation (no. 1250/18 to T.K.) and the Center for Interdisciplinary Data Science Research (to T.K., Y.D. and B.G.). N.L. was supported by CIDR Data Science and Leibniz fellowships. Y.D. holds the Walter and Greta Stiel Chair and Research Grant in Heart Studies.

Author information

Author notes

1. Einat Carmon

   Present address: Department of Surgery, Samson Assuta Ashdod University Hospital, Ashdod, Israel

2. Alex Aravanis

   Present address: Illumina, Inc., San Diego, CA, USA

3. These authors contributed equally: Netanel Loyfer, Judith Magenheim, Ayelet Peretz

Authors and Affiliations

1. School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

Netanel Loyfer, Sapir Shabi-Porat & Tommy Kaplan

2. Department of Developmental Biology and Cancer Research, Institute for Medical Research Israel-Canada, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

   Judith Magenheim, Ayelet Peretz, Agnes Klochendler, Ilana Fox-Fisher, Merav Hecht, Tsuria Pelet, Joshua Moss, Zeina Drawshy, Ruth Shemer, Yuval Dor & Tommy Kaplan

3. GRAIL, Inc., Menlo Park, CA, USA

   Gordon Cann, Joerg Bredno, Hamed Amini, Patriss Moradi, Sudharani Nagaraju, Alex Aravanis, Oliver Venn & Arash Jamshidi

4. Sharett Institute of Oncology, Hadassah Hebrew University Medical Center, Jerusalem, Israel

   Joshua Moss

5. Department of Obstetrics and Gynecology, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

   Dvora Bauman, David Shveiky, Shay Porat & Uri Dior

6. Department of Orthopedics, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

   Gurion Rivkin & Omer Or

7. Department of Otolaryngology, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

   Nir Hirshoren

8. Department of General Surgery, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

Einat Carmon, Abed Khalaileh, Gideon Zamir, Ronit Grinbaum, Machmud Abu Gazala, Ido Mizrahi & Noam Shussman

9. Surgery Division, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

Alon Pikarsky

10. Department of Cardiothoracic Surgery, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

Amit Korach, Ori Wald, Uzi Izhar & Eldad Erez

11. Department of Urology, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

Vladimir Yutkin

12. Department of Vascular Surgery, Shaare Zedek Medical Center, Jerusalem, Israel

Yaacov Samet

13. Department of Endocrinology and Metabolism, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

Devorah Rotnemer Golinkin & Benjamin Glaser

14. Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden

Kirsty L. Spalding

15. Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden

Henrik Druid

16. Department of Forensic Medicine, The National Board of Forensic Medicine, Stockholm, Sweden

    Henrik Druid

17. Department of Medicine (H7) and Karolinska University Hospital, Karolinska Institutet, Stockholm, Sweden

    Peter Arner

18. Department of Surgery and the Clinical Islet Transplant Program, University of Alberta, Edmonton, Alberta, Canada

    A. M. James Shapiro

19. Papé Family Pediatric Research Institute, Oregon Health & Science University, Portland, OR, USA

    Markus Grompe

Contributions

A.A., G.C., R.S., B.G., T.K. and Y.D. conceived and initiated the project and designed the experiments. D.B., D.S., S.P., U.D., G.R., O.O., N.H., E.C., A. Pikarsky, A. Khalaileh, G.Z., R.G., M.A.G., I.M., N.S., A. Korach, O.W., U.I., E.E., V.Y., Y.S., D.R.G., K.L.S., H.D., P.A., A.M.J.S. and M.G. provided materials. J. Magenheim, A. Peretz, A. Klochendler, I.F.-F., M.H., T.P. and Z.D. performed experiments. G.C., J.B., H.A., P.M., S.N., O.V. and A.J. conducted sequencing and analysis. N.L., S.S.-P., J. Moss and T.K. developed and performed computational analyses. N.L., Y.D. and T.K. wrote the manuscript.

Corresponding authors

Correspondence to Yuval Dor, Benjamin Glaser or Tommy Kaplan.

Ethics declarations

Competing interests

This work was supported by GRAIL, Inc. G.C., J.B., A.A., O.V. and A.J. are employees, shareholders and/or founders at GRAIL, Inc. J.M., J.M., I.F.-F., R.S., Y.D., B.G. and T.K. have filed patents on cfDNA analysis technology. The remaining authors declare no competing interests.

Peer review

Peer review information

Nature thanks Matteo Pellegrini and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 A human whole-genome DNA methylation atlas of healthy cell types.

205 healthy samples were obtained from adult humans, isolated and deeply sequenced (WGBS, mean depth ≥30x), to form a comprehensive human cell-type-specific methylation atlas.

Extended Data Fig. 2 Sample preparation and purity.

(**A**) Fresh tissue was obtained at surgery and dissociated (optimized per tissue type), then incubated with antibodies, and FACS-sorted. Sorted cells were analysed using qRT-PCR for key cell-type-specific genes, or targeted PCR for cell-type-specific DNA methylation markers.DNA methylation was

also analysed using whole-genome bisulfite sequencing. (**B**) Example of FACS sorting for pancreatic endocrine cell types. Left panel: staining for the beta cell marker C-peptide (x-axis) versus alpha cell marker glucagon (y-axis). Note that no double positive cells are observed. Centre panel: staining for c-peptide (x-axis) versus delta cell marker somatostatin (y-axis). Right panel: unstained control (only fluorescent secondary antibodies added, no primary antibodies). (**C**) Fragment-level validation of sample purity using targeted PCR. Cell-type-specific markers were designed using pre-existing 450K data, covering 4–7 several neighbouring CpGs. Shown is the percentage of unmethylated molecules in each cell type (including endothelial cells and leukocytes). Colour gradient fades from fully unmethylated molecules (allT), through those unmethylated in all but one CpG (allT-1), etc. Amplicon locations are reported in hg19, for acinar cells, alpha, beta, delta, duct, and endothelial markers (from left to right). (**D**) Fragment-level validation of the same locations, using the atlas WGBS data. Y-axis marks the percentage of unmethylated fragments (with ≥4 CpGs). As these markers show, approximately 90% of molecules in that target cell type are unmethylated, compared with less than 5% in other cell types, thus emphasizing the purity of the DNA methylation atlas using a set of independently selected DMRs.

Extended Data Fig. 3 Purity estimation for pancreas, lung, heart, liver, breast, and GI using atlas markers.

The percent of unmethylated fragments (y-axis) among fragments of ≥4 CpGs from selected differentially methylation markers could serve as an (under-) estimate of the atlas purity. Here we show one such marker for each cell type, selected from the top 25 markers, and use fragment-level analysis to demonstrate the purity in the target cell type compared to other cell types from the same tissue or environment. (**A**) Pancreas. (**B**) Lung. (**C**) Heart. (**D**) Liver. (**E**) Breast. (**F**) GI tract. For most cell types, 90% of the molecules in the target cell types are unmethylated, compared with less than 5% of other types. This is an under-estimation, as some heterogeneity could occur in each cell type, reflecting stochastic noise, cellular states, age, or environmental changes.

Extended Data Fig. 4 Biological replicates of the same cell type, from different individuals show a surprisingly low rate of differentially methylated blocks.

We focused on 37 cellular subtypes with n≥3 replicates (e.g. endothelial cells from a specific tissue) and measured the average percentage of methylation blocks (≥3 CpGs) that differ in their methylation by 50% (absolute delta beta), across replicates (shown as Y-axis). Nearly all cellular subtypes (36/37) differ by ≤0.5% of blocks suggesting a very high degree of conservation among replicates. Dotted red line marks the average number of differential blocks between two random samples of different cell types (4.9%). Box plots mark median and interquartile range (IQR), with 1.5*IQR whiskers.

Extended Data Fig. 5 Unsupervised agglomerative clustering of human methylomes.

(**A**) Same as Fig. 2, coloured by developmental lineage from germ layers, including endoderm (green), mesoderm (blue), and ectoderm (red). (**B**) same as Fig. 2, for Roadmap Epigenomics DNA methylation atlas.

Extended Data Fig. 6 Marker specificity across 953 cell-type-specific markers.

For every cell type (row), we plot each of the top 25 markers (shown as boxes). For each marker, we compare the percentage of unmethylated fragments (≥3 CpGs) in the target samples (blue dots) versus their percentage in background samples (golden dots). Blue and golden bars plot the average proportion across all target and background samples, respectively.

Extended Data Fig. 7 Markers of putative enhancers in other atlas cell types.

Including top 250 unmethylated markers for B cells (top left), hepatocytes (top right), gastric/small intestine/colon epithelium (bottom left), and

cardiomyocytes/heart fibroblasts (bottom right). Grey lines mark the same ChIP-seq/ATAC/DNase/chromHMM signal, averaged across all 11,371 unmethylated markers (top 250 per cell type).

Extended Data Fig. 8 Fragment-level analysis of *in silico* mixes.

Shown are *in silico* simulations for four cell types, which are computationally mixed at various proportions with a plasma-like mixture of 90% leukocytes, 7.5% vascular endothelial cells, and 2.5% hepatocytes. Each mixture was analysed using our atlas (red), and compared to Moss et al. (grey). Box plots show average contribution in 10 simulations, with 1 SD error bars.

Extended Data Fig. 9 Specialized DNA methylation atlases for the analysis of (A) Illumina BeadChip 450K array, (B) EPIC array, (C) RRBS sequencing data, (D) Illumina TruSeq EPIC capture panel, (E) Roche SeqCapEpi capture panel, and (F) Agilent SureSelectXT capture panel

.

Extended Data Fig. 10 Deconvolution of previously published 450K DNA methylation array data.

(**A**) Deconvolution of pancreatic islet methylomes. Methylation arrays from 53 male and 34 female non-diabetic donors[48] were analysed atlas methylomes, revealing detailed cellular composition including previously uncharacterized alpha and delta cells. No statisticaly significant sex differences in cellular composition were observed. (**B**) Analysis of 865 pulmonary methylomes from TCGA[49]. WGBS-based markers for lung alveolar epithelium and lung bronchial epithelium cells reveal differential cell populations in 443 LUAD, 11 SCLC, 337 LUSC, 32 normal adjacent (LUAD), and 42 normal adjacent (LUSC) lung methylomes. Note that only alveolar cell DNA is identified in lung adenocarcinomas, while small cell lung cancer and squamous cell carcinomas contain also bronchial DNA, consistent with the presumed cellular origins of each type of lung cancer.

Note that epithelial cells are a minority in both lung adenocarcinoma and normal lungs. This is probably due to the abundance of stromal cells in bulk preparations of either normal lungs or lung cancers. (**C**) DNA methylation from 721 cancerous and 97 normal breast biopsies from TCGA. WGBS-based markers for breast luminal and basal epithelial cells were used to study the cellular composition in TCGA[50], which were classified into five subtypes using PAM50, a 50-gene expression-based classification[66]. Different cell composition is observed for normal-like, basal-like, luminal A, luminal B, and Her2-enriched PAM50 subtypes, compared to healthy breast biopsies. The low fraction of breast basal cells in breast cancer is likely to result from the abundance of non-epithelial cells in both the normal breast and breast cancer. Box plots mark median and interquartile range (IQR), with 1.5*IQR whiskers.

Supplementary information

Supplementary Information

Legends for Supplementary Tables; cell sorting: FACS plots and RNA/methylation enrichments; genome-wide segmentation to methylation blocks; selection of type-specific markers; UXM fragment-level deconvolution algorithm; purity estimations: proportion of unmethylated fragments in various samples; analysis of cfDNA composition in patients with COVID-19; dissociation protocols; appendix: RBC lysis buffer (X10); Supplementary Table 18: antibodies for flow cytometry; and Supplementary Table 19: reagents used.

Reporting Summary

Peer Review File

Supplementary Tables

This file contains Supplementary Tables 1–17 (see main Supplementary Information PDF for legends).

Supplementary Dataset 1

Genome-wide set of unmethylated regions per cell type, annotated. Zip file contains 39 bed files, each with all genomic regions (blocks of at least four CpGs in which at least 85% of sequenced fragments are unmethylated in at least 85% of covered CpGs). Bed file information includes position (chr, from, to, CpG_from, CpG_to, hg19), regional annotation by HOMER and neighbouring gene, and annotations of when a region is 1 kb or less from a transcriptional start site, overlapping a CpG island, chromHMM annotation in that cell type or similar, ChIP–seq peak annotations for H3K4me3 (promoter), H3K27ac (gene regulation), H3K4me1 (enhancer), H3K27me3, overlapping ATAC–seq peak, presence of CTCF binding site and CTCF ChIP–seq peak.

Supplementary Dataset 2

Genome-wide set of unmethylated regions per cell type, overlapping H3K27ac, but not J3K4me3, ChIP–seq peaks, annotated. Zip file containing bed files, listing for each cell type a set of putative enhancers defined as the subset of hypomethylated regions (Extended Dataset 1) marked by active gene regulation chromatin marks (H3K27ac peaks) but lacking promoter marks (H3K4me3 peaks).

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Loyfer, N., Magenheim, J., Peretz, A. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023). https://doi.org/10.1038/s41586-022-05580-6

- Received: 30 November 2021
- Accepted: 18 November 2022
- Published: 04 January 2023
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05580-6

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

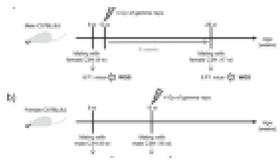
Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

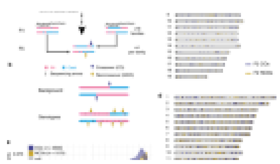
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05580-6



Characteristics of induced mutations in offspring derived from irradiated mouse spermatogonia and mature oocytes

08 January 2020

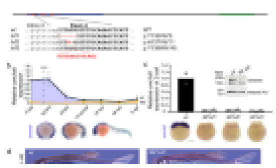
Yasunari Satoh, Jun-ichi Asakawa, … Arikuni Uchimura



A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination

29 August 2019

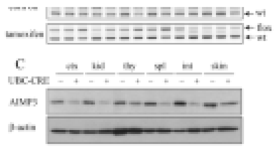
Ran Li, Emmanuelle Bitoun, … Simon R. Myers



HOX epimutations driven by maternal SMCHD1/LRIF1 haploinsufficiency trigger homeotic transformations in genetically wildtype offspring

23 June 2022

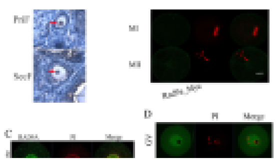
Shifeng Xue, Thanh Thao Nguyen Ly, … Bruno Reversade



AIMP3 Deletion Induces Acute Radiation Syndrome-like Phenotype in Mice

09 October 2018

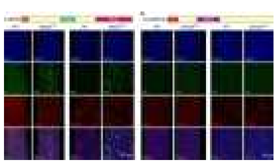
Doyeun Kim, Sunmi Kim, … Sunghoon Kim



Rad9a is involved in chromatin decondensation and post-zygotic embryo development in mice

28 August 2018

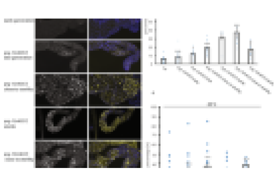
Lin Huang, Tie-Gang Meng, … Qing-Yuan Sun



Defective germline reprogramming rewires the spermatogonial transcriptome

30 April 2018

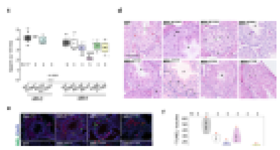
Lina Vasiliauskaitė, Rebecca V. Berrens, … Dónal O’Carroll



[Germ granule dysfunction is a hallmark and mirror of Piwi mutant sterility]

03 March 2021

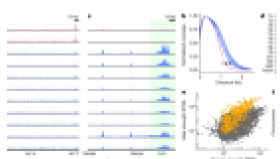
Maya Spichal, Bree Heestand, … Shawn Ahmed



[Human MLH1/3 variants causing aneuploidy, pregnancy loss, and premature reproductive aging]

18 August 2021

Priti Singh, Robert Fragoza, … John C. Schimenti



[Extensive sex differences at the initiation of genetic recombination]

05 September 2018

Kevin Brick, Sarah Thibault-Sennett, … Galina V. Petukhova

- Article
- Open Access
- [Published: 21 December 2022]

Inheritance of paternal DNA damage by histone-mediated repair restriction

- Siyao Wang ,
- David H. Meyer &
- Björn Schumacher

Nature **volume 613**, pages 365–374 (2023)

- 25k Accesses
- 1 Citations
- 368 Altmetric
- Metrics details

Abstract

How paternal exposure to ionizing radiation affects genetic inheritance and disease risk in the offspring has been a long-standing question in radiation biology. In humans, nearly 80% of transmitted mutations arise in the paternal germline[1], but the transgenerational effects of ionizing radiation exposure has remained controversial and the mechanisms are unknown. Here we show that in sex-separated *Caenorhabditis elegans* strains, paternal, but not maternal, exposure to ionizing radiation leads to transgenerational embryonic lethality. The offspring of irradiated males displayed various genome instability phenotypes, including DNA fragmentation, chromosomal rearrangement and aneuploidy. Paternal DNA double strand breaks were repaired by maternally provided error-prone polymerase theta-mediated end joining. Mechanistically, we show that depletion of an orthologue of human histone H1.0, HIS-24, or the heterochromatin protein HPL-1, could significantly reverse the transgenerational embryonic lethality. Removal of

HIS-24 or HPL-1 reduced histone 3 lysine 9 dimethylation and enabled error-free homologous recombination repair in the germline of the F_1 generation from ionizing radiation-treated P_0 males, consequently improving the viability of the F_2 generation. This work establishes the mechanistic underpinnings of the heritable consequences of paternal radiation exposure on the health of offspring, which may lead to congenital disorders and cancer in humans.

Main

Ionizing radiation induces DNA double strand breaks (DSBs) that can cause de novo mutations (DNMs) and chromosomal aberrations[2]. Somatic DNMs are highly correlated with cancer and aging[3,4]. Germline DNMs can be transmitted to offspring and are a driving force of genome evolution[5], but may also result in congenital diseases[6]. In humans, nearly 80% of DNMs[1]—including 70% of de novo structural variants[7,8]—originate from the paternal germline. Increasing evidence suggests that the paternal germline DNMs are associated with various congenital disorders, schizophrenia, autism and reproduction defects[1,9,10]. The correlation between parental exposure to ionizing radiation and genetic effects in the progeny has nonetheless remained uncertain[11,12,13]. Contradictory observations have been obtained from children born in the vicinity of nuclear power plants[14,15,16]. Studies on the progeny of the clean-up workers from the accident at Chernobyl and of survivors of the atomic bombings at Hiroshima and Nagasaki showed no evidence of a transgenerational effect of parental radiation exposure[17,18,19,20]. However, epidemiologic studies in humans are typically limited by the sample size, the mixed population, and records of radiation doses. By contrast, paternal exposure to chemotherapy before conception has recently been correlated with germline hypermutation[21], which is associated with rare genetic diseases in the progeny[22]. This raises the question of whether specifically paternal exposure to mutagenic agents could lead to a transgenerational effect.

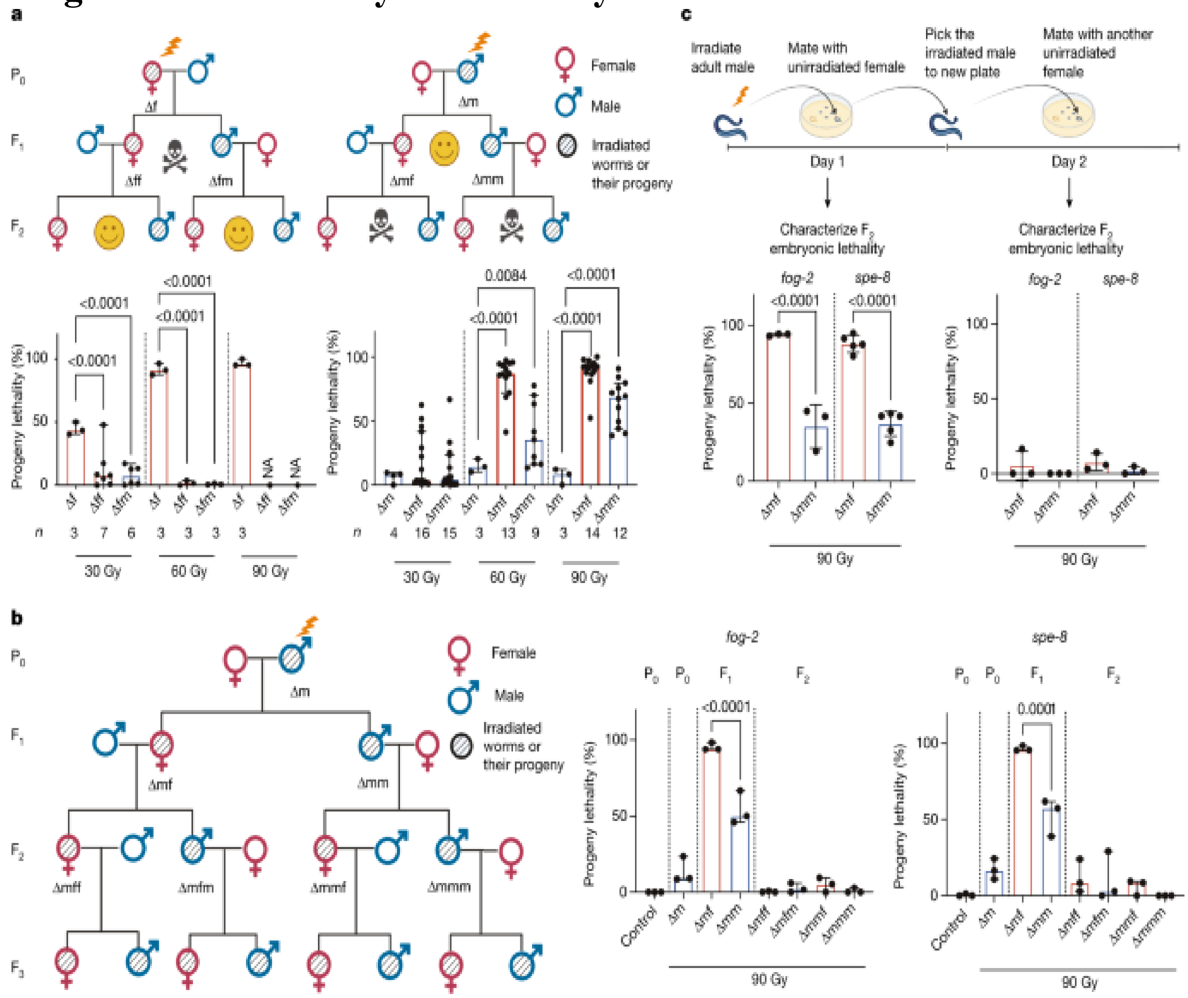
Here, we used sex-separated mutants of *C. elegans* to investigate the transgenerational effects of paternal radiation exposure. We show that

paternal exposure to ionizing radiation results in genome instability in the F_1 generation and transgenerational embryonic lethality. We determined that the paternal DNA damage is mainly repaired in the zygote through maternally provided error-prone polymerase theta-mediated end joining (TMEJ), which results in chromosomal aberrations. Structural variants with such TMEJ signature microhomology are also present in natural *C. elegans* variants and in paternal germline mutations in humans, suggesting a conserved mechanism. Removing the histone H1 HIS-24 or the heterochromatin protein HPL-1 could reverse the transgenerational embryonic lethality of paternal exposure to ionizing radiation by activating the error-free homologous recombination repair (HRR) pathway. Our work provides the mechanistic underpinnings of the transgenerational genetic and epigenetic effect of paternal exposure to ionizing radiation.

Ionizing radiation induces a transgenerational effect

To investigate the transgenerational consequences of genotoxic stress, we exposed *C. elegans* to ionizing radiation and assessed the consequences in the subsequent generations. Wild-type *C. elegans* has two sex forms: the self-fertilizing hermaphrodites and males. To distinguish the effect of DNA damage in male and female gametes, weused the feminized mutant *fog-2*, which has a dioecious reproductive system: somatic hermaphrodites that only generate a female germline and males with a normal male germline. Immediately after ionizing radiation treatment, females (Δf) were mated with unirradiated males and the viability of the progeny was characterized 24 h later. As expected, ionizing radiation treatment led to a dose-dependent embryonic lethality (Fig. 1a; full statistical analyses for all results are shown in Supplementary Table 1), which is attributed to the excessive unrepaired DNA DSBs in oocytes. When we crossed the surviving female (Δff) and male (Δfm) F_1 worms with healthy worms of the opposite sex, the F_2 generation showed nearly no embryonic lethality (Fig. 1a). This observation indicates that the DNA DSBs carried in the oocytes can either be properly repaired or eliminated, thereby preventing any transgenerational effect from the maternal DNA damage.

Fig. 1: Paternal exposure to ionizing radiation leads to transgenerational embryonic lethality.



a, Pedigrees of the maternal and paternal exposure to different doses of ionizing radiation and transgenerational characterization. Left, maternal exposure to ionizing radiation (P_0 Δf) leads to intergenerational embryonic lethality in the F_1 generation. Except for the group treated with 90 Gy, which has no surviving progeny developing to adulthood, the surviving *fog-2* F_1 (Δff and Δfm) crossed with non-irradiated opposite sexes recovered the embryonic lethality to basal level. *n* indicates the number of biological replicates; each replicate includes three *fog-2* females and three *fog-2* males. Right, paternal ionizing radiation exposure (P_0 Δm) leads to a mild increase in embryonic lethality in the F_1 generation, and the progeny of the surviving F_1 (Δmf and Δmm) show a transgenerational embryonic lethality in the F_2 generation. *n* indicates the number of biological replicates. Data are median

± 95% confidence interval; *P* values are shown. Throughout the figures, red bars indicate females and blue bars indicate males. NA, not applicable. **b**, Pedigrees of paternal ionizing radiation exposure (90 Gy) and the transgenerational lethality characterization for three consequent generations. $n = 3$ biological replicates, each replicate includes 3 females with 3 males. Data are median ± 95% confidence interval; *P* values are shown. **c**, Freshly irradiated P_0 adult males and males two days after ionizing radiation exposure show different transgenerational effects. *fog-2*: $n = 3$ biological replicates; *spe-8*: $n = 5$ biological replicates; each replicate includes 3 females and 3 males. Data are median ± 95% confidence interval; *P* values are shown. Generalized linear model (GLM) with logit link function and Tukey's multiple comparisons were used for proportional data, and one-way (**a**,**b**) or two-way (**c**) ANOVA with arcsine transformation was also used to confirm the statistical results (Supplementary Table 1).

We next tested whether ionizing radiation treatment in males (Δm) could affect subsequent generations. Immediately after ionizing radiation exposure, males were mated with untreated females. The progeny lethality of Δm was relatively low (Fig. 1a), indicating that the paternal DNA damage had either been repaired or escaped from the checkpoint surveillance during the embryogenesis of the F_1 generation. Unexpectedly, after crossing the F_1 worms with healthy worms of the opposite sex, the embryonic lethality of the F_2 generation was markedly increased (Fig. 1a). Notably, this increase in embryonic lethality is significantly different between the progeny of female F_1 (Δmf) and the progeny of male F_1 (Δmm) worms (Fig. 1a). These observations establish that male exposure to ionizing radiation can lead to transgenerational lethality in the F_2 generation.

Subsequent mating of the surviving progeny from the F_2 generation with the healthy opposite sex restored the viability of the progeny (Fig. 1b), indicating that the sperm-inherited detrimental effect can be largely eliminated within two generations of intercrossing with healthy worms. We also verified this observation with an independent feminized mutant, *spe-8* (Fig. 1b).

Probably owing to the limited DNA repair capacity within a highly condensed chromosomal architecture, mature sperm is more vulnerable than

spermatocytes to DNA damage[23]. As in humans, spermatogenesis in male *C. elegans* is continuously ongoing in adults, and in one nematode mating about two-thirds of the mature sperm are ejaculated. To test whether the mature sperm might be responsible for the transgenerational lethality, we first collected the progeny derived from freshly irradiated *fog-2* or *spe-8* males to assess the effect of irradiation on mature sperm. On the second day, we mated those irradiated males with untreated females to test the effect of irradiation on spermatogenesis before maturation. Only the progeny derived from the freshly irradiated males led to transgenerational lethality (Fig. 1c), indicating that only mature spermatozoa were responsible for the ionizing radiation-induced transgenerational lethality.

We next tested whether the transgenerational lethality also occurs in wild-type *C. elegans* hermaphrodites. Spermatogenesis is initiated at the late L3 larval stage and ceases by the late L4 larval stage, where the germline switches to oogenesis while the mature sperm is stored in the spermatheca. Only irradiation of wild-type hermaphrodites at the late L4 stage but not at late L3 or young L4 led to transgenerational lethality in the F_2 and F_3 generations (Extended Data Fig. 1a), validating that ionizing radiation exposure of mature sperm in hermaphrodite wild-type worms could lead to transgenerational lethality.

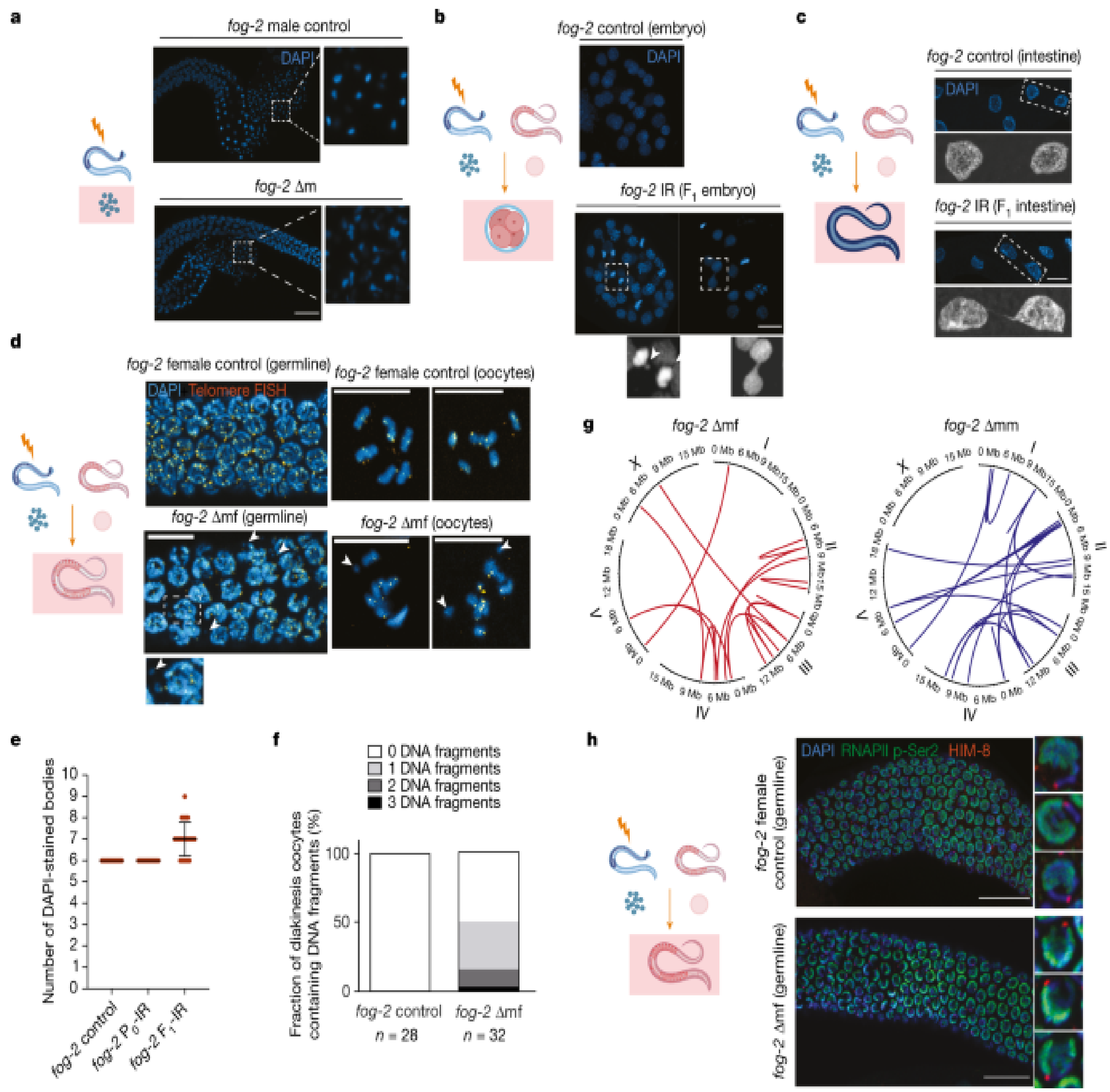
In contrast to *fog-2* mutants, wild-type hermaphrodites self-fertilize and might thus further perpetuate the unstable genomes. Indeed, we observed that the transgenerational lethality was maintained over at least six generations (Extended Data Fig. 1b). These results suggest that the recovery of viability in the F_3 generation of *fog-2* mutants is mainly owing to the dilution of the transgenerational effect by outcrossing the damaged genomes three times with the unirradiated opposite sex.

Paternally induced genome instability

To address whether the paternal DNA damage causes heritable changes, we tracked chromatin states from the irradiated sperm to the germ cells of the F_1 generation. Ionizing radiation treatment led to morphological changes in the sperm of *fog-2* Δm with a ‘comet-like’ conformation (Fig. 2a and Extended Data Fig. 2a). After fertilization, the embryos showed various chromosomal

instability phenotypes, including the formation of chromosomal bridges and chromosome lagging (Fig. 2b and Extended Data Fig. 2b). The formation of chromosomal bridges indicated that the dividing embryonic cells entered a ‘breakage–fusion–bridge’ cycle[24]. Despite these chromosomal aberrations, the majority of embryos developed into phenotypically healthy adults, suggesting that the DSBs in the sperm bypass the surveillance of mitotic checkpoints during early embryogenesis. Unlike the paternally irradiated embryos, DNA damage induced by maternal exposure to radiation mainly caused fragmented DNA in the F_1 embryos (Extended Data Fig. 2b,c). Such DNA fragmentation could further trigger mitotic catastrophe, accounting for the high embryonic lethality in the F_1 generation.

Fig. 2: Paternal exposure to ionizing radiation leads to DNA fragmentation, chromosomal rearrangement and aneuploidy in the F_1 generation.



a, Left, male germline DAPI staining. Right, magnified mature sperm. Scale bar, 20 μm. **b**, Embryonic DAPI staining. Bottom, chromosomal lagging and bridging. Scale bar, 10 μm. **c**, Intestinal DAPI staining. Bottom, magnified view of two representative cells. Scale bar, 10 μm. **d**, Telomere fluorescence in situ hybridization (FISH) and DAPI staining in the germline. Bottom, magnified view of fragmented DNA. Right, late diakinesis oocytes. Arrowheads indicate DNA fragmentation. Scale bars, 10 μm. Experiments in **a**–**c** were repeated three times with similar observations; experiment in **d** was repeated two times with similar observations. **e**, Quantification of DAPI-stained bodies in late diakinesis oocytes. Data are mean ± s.d., $n = 25$ oocytes. **f**, Quantification of DNA fragmentation (without telomere FISH

signal) per late diakinesis oocyte. *n* represents the number of oocytes. **g**, Representative circos plots showing chromosomal translocations (inter- or intra-chromosomal fusions) in Δmf and Δmm. All translocations are listed in Supplementary Table 3. **h**, RNAPll p-Ser2, HIM-8 and DAPI staining in dissected germline. Right, three representative nuclei. Scale bar, 20 μm. This experiment was repeated three times with similar observations. IR indicates exposure to ionizing radiation. Drawings illustrate the (**a**) sperm, (**b**) F_1 embryos, (**c**) F_1 somatic tissues and (**d**, **h**) F_1 germlines of irradiated males that are investigated in the respective panels.

We next characterized the chromosomal states in the adult *fog-2* F_1 offspring with paternally inherited DNA damage. Intestinal nuclei that undergo a set of programmed endoreduplications during larval development displayed chromosomal bridges in the *fog-2* F_1 offspring of *fog-2* Δm (Fig. 2c and Extended Data Fig. 2d), indicating the persistence of DNA breaks from embryogenesis until adulthood. We observed similar chromosomal bridges in the hermaphrodite F_1 offspring from irradiated L4 larvae (Extended Data Fig. 2e). The intestinal chromosomal bridges were absent in the F_1 generation of maternal radiation exposure, indicating that the genome stability of surviving F_1 progeny from irradiated mothers was restored (Extended Data Fig. 2d). This observation was also confirmed in *spe-8* mutants (Extended Data Fig. 2d).

We also observed a developmental growth delay in the Δmf and Δmm of both *fog-2* and *spe-8* mutants (Extended Data Fig. 2f), suggesting that paternal exposure to ionizing radiation can lead to phenotypic consequences in the progeny.

In the pachytene zone of the germline of *fog-2* Δmf worms, we observed DNA fragments (Fig. 2d, left) that were devoid of telomere sequence (Fig. 2d, left). Instead of the normal six highly condensed and discrete oocyte bivalents, meiotic diakinesis cells of the *fog-2* Δmf germlines contained a larger number of DAPI-stained bodies that were irregularly sized, with some carrying randomly allocated telomere sequences (Fig. 2d, right, e,f). The chromosomal rearrangement, fragmentation and aneuploidy in the F_1 generation might also provide an explanation for why the transgenerational

effect is limited to three generations in the *fog-2* mutants but lasts more than six generations in hermaphrodite wild-type worms, as out-crossing with normal haploid gametes could greatly rescue the inheritance of aneuploidy.

The chromosomal aberrations prompted us to test whether the apoptotic DNA damage checkpoint was activated in the F_1 germline. As the male germline lacks this apoptotic response[25], we quantified the apoptotic corpses in the germlines of control, Δf and Δmf worms. Ionizing radiation induced a significant increase of apoptosis in the germlines of Δf worms; however, only a mild increase of apoptosis was detected in the germline of Δmf (Extended Data Fig. 2g) indicating that the paternally inherited genome instability does not trigger the DNA damage checkpoint in the F_1 germline.

To detect de novo genomic aberrations in the *fog-2* F_1 generation, we performed whole-genome sequencing (WGS) of single P_0 and their direct F_1 progeny. We detected between 3 and 21 de novo translocations (inter- or intra-chromosomal fusions) in each female (Δmf) and between 2 and 28 de novo translocations in each male (Δmm) F_1 individual. In total, we mapped 154 translocations in the genomes of *fog-2* Δmf and *fog-2* Δmm worms compared with the parental generation (Extended Data Fig. 2h and Supplementary Table 2), whereas tandem duplications ($n = 1$) and insertions ($n = 1$) were rare (Supplementary Table 3). All de novo translocations in the F_1 worms were heterozygous, consistent with only paternally inherited genomic aberrations. The genomes of female and male progeny showed an overall similar number of autosomal translocations. The paternal X chromosome is inherited in *fog-2* Δmf but not in *fog-2* Δmm worms, which carry only a maternal X chromosome. Consistently, only the *fog-2* Δmf worms bear X chromosomal translocations (Fig. 2g, Extended Data Fig. 2i,j and Supplementary Table 3), indicating that the translocations occur only in the paternal chromosomes.

The X chromosomal translocations might be responsible for the particularly high progeny lethality in the *fog-2* Δmf, as the integrity of the X chromosome is essential for meiotic sex chromosomal inactivation (MSCI). When the continuity of the X chromosome is lost, X chromosomes cannot enter the MSCI transcriptional inactivation state, resulting in high progeny lethality[26]. To assess whether MSCI is disrupted, we stained a transcription

active mark, Ser2 phosphorylation (p-Ser2) of RNA polymerase II (RNAPII). In the pachytene region of the germline, X chromosomes (labelled with HIM-8 antibody, marking the X chromosome pairing centre) in the untreated *fog*-2 females were devoid of RNAPII p-Ser2, whereas in the *fog-2* Δmf, RNAPII p-Ser2 was evenly distributed on both autosomes and X chromosomes (Fig. 2h). Consistently, gene set enrichment analysis (GSEA) for chromosomal gene distributions indicates a strong upregulation of X chromosomal proteins in the F_1 generation of irradiated wild-type worms (Extended Data Fig. 3 and Supplementary Table 4), whereas no other chromosome was significantly enriched (Extended Data Fig. 3 and Supplementary Table 4). These observations suggest that the X chromosome rearrangement in the germline of *fog-2* Δmf disrupts the MSCI process, which together with chromosomal translocations and fragmentation on autosomes, results in significantly higher progeny lethality compared with the *fog-2* Δmm.

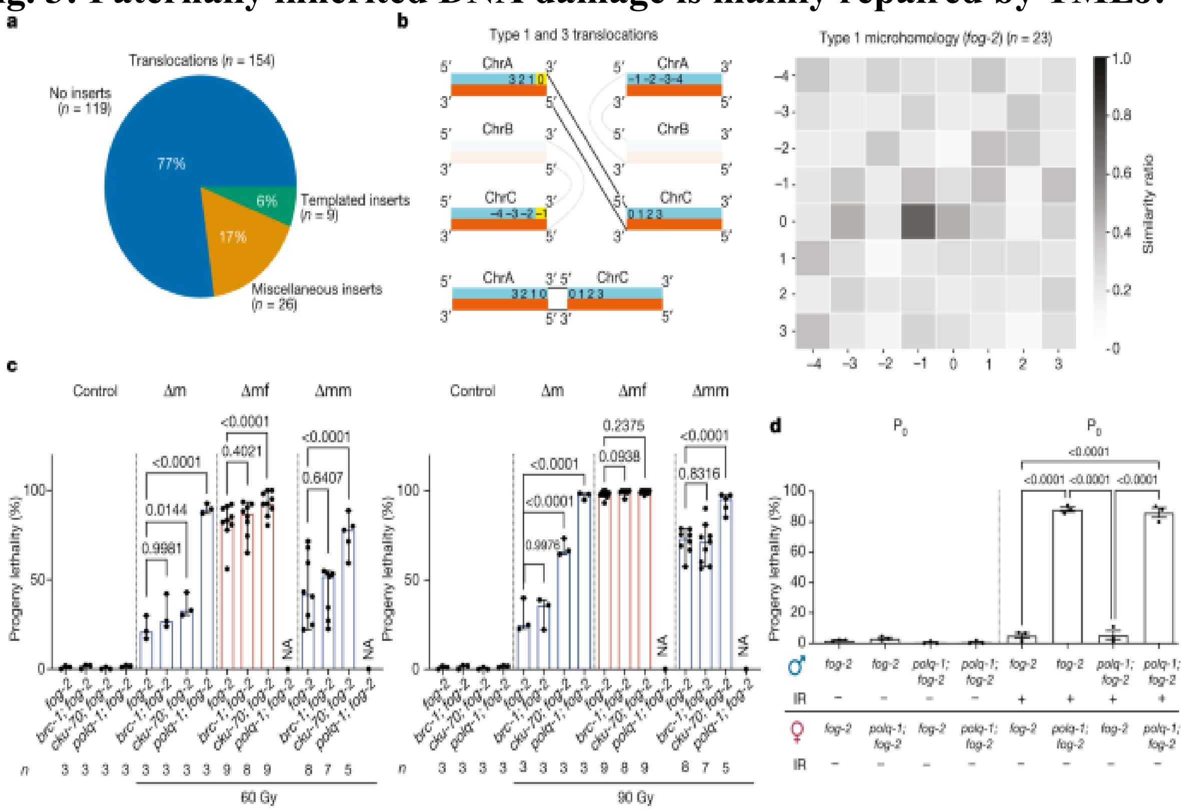
In addition to *fog-2* mutants, we also performed WGS of single hermaphrodite wild-type F_1 worms and their P_0 parents. Unlike the dioecious *fog-2* system, the self-fertilized hermaphrodites carried roughly twice the number of heterozygous translocations in the genome of their F_1 generation (Extended Data Figs. 4a and 5). This could be explained by two possibilities: (1) the sperm of the hermaphrodites is more vulnerable than the sperm produced in the male worms; (2) in contrast to self-fertilizing wild-type hermaphrodites, mating of the *fog-2* males with healthy females providing a set of undamaged maternal chromosomes could greatly increase genome stability. Similarly to the above result, we also observed aneuploidy phenotypes and de-silencing of the X chromosome in the F_1 of irradiated hermaphrodite wild-type worms (Extended Data Fig. 4b,c).

Maternal TMEJ repairs paternal DNA

DSBs are mainly repaired by either the error-free HRR or the error-prone non-homologous end joining (NHEJ). In cells deficient in NHEJ or HRR, error-prone TMEJ is crucial for viability[27]. TMEJ involves a microhomology in DNA flanking the breaks and uses DNA close to the breakpoints as a template for de novo synthesis[28]. We analysed the sequence

context of the pooled 154 de novo translocations identified in the F_1 generation of *fog-2* worms. Whereas most translocations showed no insertions in between their breakpoints, around 23% contained at least 1 inserted base pair, and 6% had an insertion of at least 3 bp whose template could be found within ±25 bp around the breakpoints, thereby matching the TMEJ pattern defined in *C. elegans*[28] (Fig. 3a and Extended Data Fig. 6a). Wild-type samples showed a similar pattern, with 11% templated inserts (Extended Data Fig. 6b,c). This suggests that translocation-inducing error-prone DSB repair could use DNA close to the breakpoints as a template for de novo synthesis preceding end joining.

Fig. 3: Paternally inherited DNA damage is mainly repaired by TMEJ.



a, The distribution of translocation footprints in *fog-2* F_1 adults with paternal exposure to 90 Gy ionizing radiation. Templated inserts are insertions of ≥3 bp in between the fusion sites that have a matching sequence within ±25 bp around one of the 2 break points. Miscellaneous insertions are insertions <3 bp or insertions with no matching sequence within ±25 bp around the breakpoints. *n*, number of translocations. **b**, Left, schematic illustration of the type 1 translocations. The sense strand is in blue, and the antisense

strand is in red. Numbers indicate the distance of the nucleotide to the breakpoint; yellow boxes indicate positions that show microhomology. Right, heat map representing the sum of all type 1 translocation maps derived from Δmf and Δmm ($n = 23$ translocations), excluding translocations with a templated insertion. Darker shades indicate higher sequence similarity between the corresponding bases. Numbers along the *x*- and *y*-axes correspond to those on the left. *P* values are provided in Supplementary Table 1. **c**, Progeny lethality of indicated strains. *n* represents the number of biological replicates. Data are median ± 95% confidence interval. *P* values are shown. **d**, Progeny lethality of different parental combinations for *polq-1* worms with and without exposure to 90 Gy ionizing radiation. $n = 3$ biological replicates. Data are median ± 95% confidence interval. *P* values are shown. GLM with logit link function and Tukey multiple comparisons were used for proportional data, and one-way (**c**) or two-way (**d**) ANOVA with arcsine transformation was used to confirm the statistical results. Full statistical analyses are provided in Supplementary Table 1.

Given the TMEJ signature, we next tested whether microhomology was present in the sequence around the breakpoints of the de novo translocations. We divided the translocations into four types on the basis of their sequence orientation and identified a microhomology in each of the types. Types 1 and 3 are shown in Fig. 3b and types 2 and 4 are shown in Extended Data Fig. 6d–f. Types 1 and 3 annotate the same translocation—one is annotated from the 5′ direction and the other is from the 3′ direction. For clarity, we refer to both as type 1. We assessed the degree of sequence similarity in an 8-bp window around the breakpoints. The last base pair within the translocation—that is, the breakpoint—is denoted with zero, the three following bases are denoted with positive numbers, and the last four bases that are not included in the translocation are denoted with negative numbers (Fig. 3b right, and Extended Data Fig. 6e). We identified the same patterns in the genomes of wild-type F_1 worms (Extended Data Fig. 6f). Finally, we investigated the base composition around the breakpoints (Extended Data Fig. 6g). The sequence context of DNA breakpoints is thus reminiscent of the outcomes of error-prone TMEJ.

To corroborate the repair pathway usage, we crossed the mutants of the major DSB repair pathways with *fog-2*: HRR is inactivated in *brc-1* mutants, NHEJ is inactivated in *cku-70* mutants, and TMEJ is inactivated in *polq-1*

mutants. HRR defective *brc-1; fog-2* mutants did not further increase the progeny lethality of the P_0 and F_1 generations (Fig. 3c), indicating that the paternally inherited DNA damage does not engage the error-free HRR repair pathway. By stark contrast, the TMEJ mutant *polq-1; fog-2* worms were hypersensitive to paternal DNA damage (Fig. 3c), and the rare F_1 survivors were sterile owing to a germline developmental defect. Depletion of the NHEJ repair protein CKU-70 exacerbated the progeny lethality in the *cku-70; fog-2* Δm and *cku-70; fog-2* Δmm compared with *fog-2* single mutants. We further confirmed these observations by using RNA-mediated interference (RNAi) to knock down these repair proteins in the *fog-2* mutant (Extended Data Fig. 6h). These results indicate that the paternal DNA damage is repaired predominantly by TMEJ and partially by NHEJ. Given that TMEJ and NHEJ are both error-prone repair pathways[27,28,29], these findings provide an explanation for the chromosomal abnormalities observed in the F_1 generation.

We next tested whether TMEJ was active in mature sperm or after fertilization with damaged sperm. We sex-separated the *polq-1* mutants and crossed each combination between male and female *fog-2* mutants proficient or deficient (*polq-1*) for TMEJ. Progeny lethality increased to nearly 100% only when *polq-1* females were mated with males treated with ionizing radiation, regardless of paternal *polq-1* status (Fig. 3d), indicating that maternally provided TMEJ is the major route for repair of mature sperm DNA.

To test whether such a transgenerational effect of TMEJ repair outcome might also occur under physiological conditions, we identified the TMEJ signature in de novo germline mutations in natural *C. elegans* populations and in human germline mutation datasets. TMEJ signatures have previously been identified in five different natural isolates of *C. elegans*[28]. Here we found a highly significant over-enrichment of 1-bp microhomology around the breakpoints of deletion sites in 540 natural isolates of the CeNDR database[30] (Extended Data Fig. 7a,f). A recent mutation accumulation experiment with wild-type and differentDNA repair mutants involving several mutagens showed a pattern with a substantial microhomology footprint[31]. Re-analysing the deletions of samples treated without any mutagen still showed a significant overrepresentation of the 1-bp

microhomology TMEJ footprint (Extended Data Fig. 7b,f). These results are consistent with a requirement for TMEJ for repairing DSBs in germ cells and might therefore be a driver of evolution in unperturbed conditions[32].

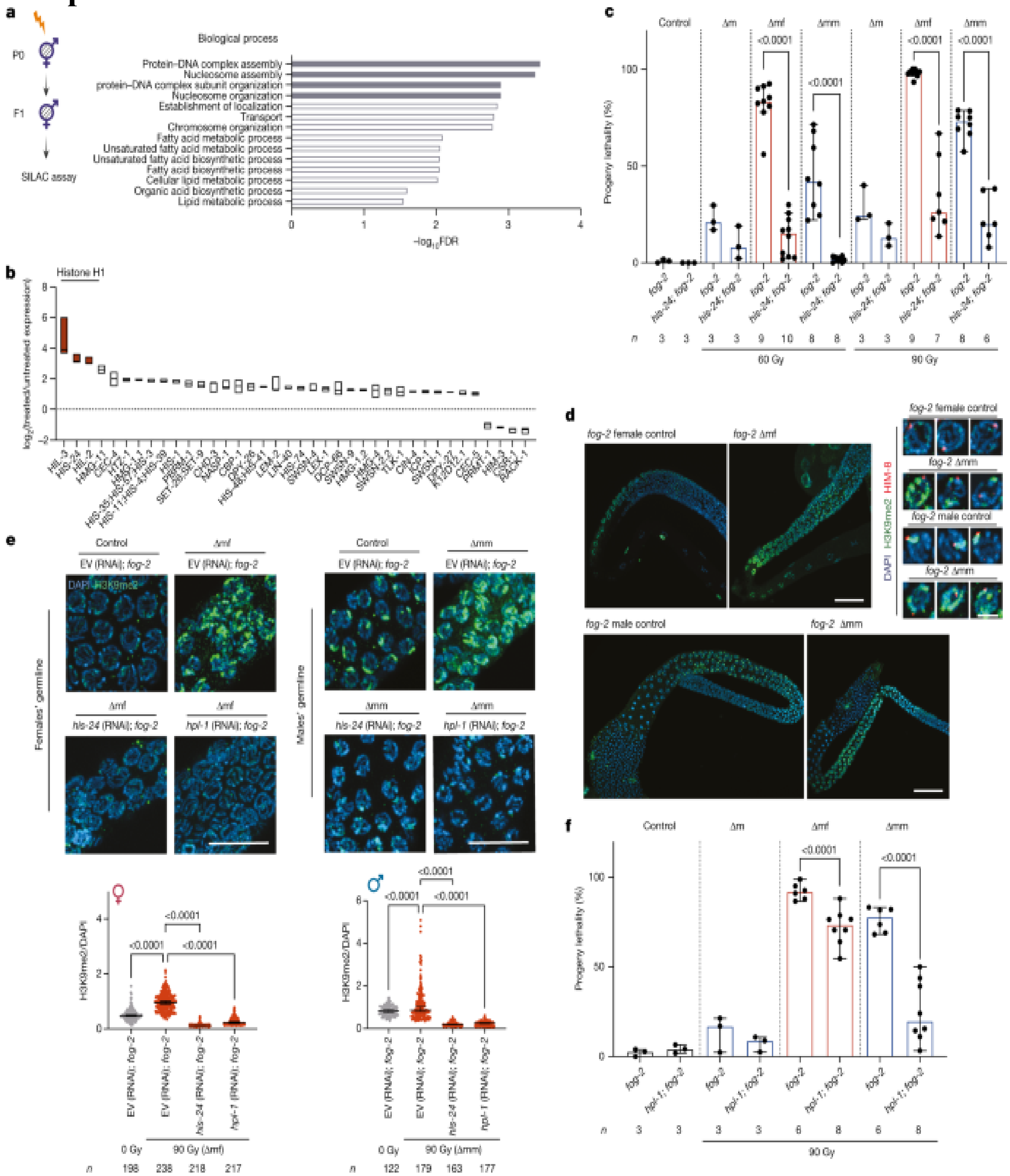
In humans, TMEJ has a 2- to 6-bp microhomology footprint[27,33]. In the 1000 Genomes Project[34], we found 2- to 6-bp microhomologies in 28% of deletions (Extended Data Fig. 7c,f). These data indicate that the TMEJ footprint is over-represented in natural human variation sites. Similarly, the de novo structural variants in the children of trios of the Polaris dataset (https://github.com/Illumina/Polaris) showed a significant over-representation of microhomology (Extended Data Fig. 7d,f). A recent gamete-of-origin analysis of de novo structural variants in 2,396 families showed that most de novo structural variants are generated in sperm and that children with a sporadic autism spectrum disorder show an increased structural variant count[8]. Moreover, microhomologous structural variants showed an increased ratio for the father as the parent of origin. Similarly, in 1,548 trios from Iceland with known gamete of origin, more de novo deletions were generated in sperm than in oocytes[35]. In our re-analysis of the published mutation calls, the deletions generated in sperm specifically showed a significant microhomology over-representation (Extended Data Fig. 7e,f). TMEJ might thus also have a role in sperm-induced mutagenesis under unchallenged conditions in *C. elegans* as well as in humans.

Histone H1 and HPL-1 trigger lethality

To identify the global changes in the proteome of the F_1 worms of parents treated with ionizing radiation, we performed stable isotope labelling with amino acids (SILAC)-based proteomics (Fig. 4a) using wild-type hermaphrodites for feasibility. The proteome changes were highly enriched in protein–DNA assembly, nucleosome assembly and nucleosome organization (Fig. 4a), with the majority of the proteins being up-regulated in the F_1 of worms treated with ionizing radiation, especially the histone proteins (Fig. 4b). The three most highly up-regulated proteins were the linker histone (H1) variants, HIL-3, HIS-24 and HIL-2. Knockdown of *hil-2* by RNAi led to larval arrest regardless of irradiation, whereas *hil-3* RNAi showed only a mild, but not significant, reduction of lethality in the F_1

progeny. Notably, RNAi knockdown of *his-24* significantly restored the embryonic viability of the wild-type F_2 generation (Extended Data Fig. 8a).

Fig. 4: Removal of histone H1 or HPL-1 alleviates lethality inherited from sperm.



a, Gene Ontology term analysis for biological processes among significantly regulated proteins from a SILAC assay for wild-type worms with or without

parental exposure to ionizing radiation. The top four enriched processes are highlighted in grey. FDR, false discovery rate. Values of $\log_{10}$FDR larger than 1.3 correspond to FDR < 0.05. **b**, Relative expression of proteins involved in chromosome- or DNA-binding processes upon parental exposure to ionizing radiation. The floating bar shows the range of values, and the centre line represents the median. *n* = 4 biological replicates. **c**, Progeny lethality characterization of the *his-24; fog-2* mutant with or without paternal ionizing radiation irradiation. *n* indicates the number of biological replicates. Data are median ± 95% confidence interval. *P* values are shown. **d**, Left, H3K9me2 (green) and DAPI (blue) staining in the germline of indicated strains. Scale bars, 20 μm. Right, three representative nuclei co-stained with the X chromosome marker HIM-8 (red). Scale bar, 3 μm. **e**, H3K9me2 (green) and DAPI (blue) staining in the germline of indicated RNAi strains with or without paternal ionizing radiation exposure. Scale bar, 10 μm. Bottom, quantification of H3K9me2 signal intensity. *n* indicates cell numbers. Data are median ± 95% confidence interval. *P* values are shown; one-way ANOVA with Bonferroni's multiple comparisons test. Experiments in **d**,**e** were repeated three times with similar observations. EV, empty vector. **f**, Progeny lethality of *fog-2* and *hpl-1; fog-2* with or without paternal ionizing radiation irradiation. *n* indicates the number of biological replicates. Data are median ± 95% confidence interval. *P* values are shown; GLM with logit link function was used for proportional data (**c**,**f**), and two-tailed *t*-test with arcsine transformation was used to confirm the statistical results. Full statistical analysis is shown in Supplementary Table 1.

Consistently, *his-24; fog-2* double mutants significantly suppressed the transgenerational embryonic lethality caused by paternal ionizing radiation exposure (Fig. 4c). However, the rescue effect of HIS-24 depletion was quickly reduced after maintaining this strain for multiple generations, implying that the function of HIS-24 can be gradually compensated by other regulators. To circumvent such an adaptation, we knocked down *his-24* in *fog-2* mutants using RNAi. The viability was significantly restored in the progeny of Δmm, whereas *his-24* RNAi was insufficient to reduce the high lethality levels in progeny of Δmf, probably because even one aberrant X chromosome structure is sufficient to prevent the MSCI, thus resulting in lethality (Extended Data Fig. 8b). These data indicate that the inherited

paternal DNA damage leads to an increase in the level of histone H1, which is responsible for the embryonic lethality in the F_2 generation.

Histone H1 is involved mainly in DNA packaging and chromosomal condensation, and functions together with heterochromatin proteins to regulate chromatin mobility, higher-order chromatin structure and transcription[36,37]. Thus the up-regulated histone H1 in the F_1 generation might be associated with an increase in heterochromatin regions. In non-irradiated *fog-2* mutants, the H3K9me2 heterochromatin mark was weakly but evenly distributed on the chromatin of late pachytene cells of the females and strongly enriched on the X chromosome in the pachytene cells of the males, as it is essential for the X chromosome silencing during meiosis in the male germline (Fig. 4d). In the germline of the F_1 worms, we found a marked increase in the level of H3K9me2 that was not specifically associated with the X chromosome, but was randomly distributed (Fig. 4d). Depletion of HIS-24 significantly abolished the increased H3K9me2 in the germlines of both *fog-2* Δmf and *fog-2* Δmm worms (Fig. 4e).

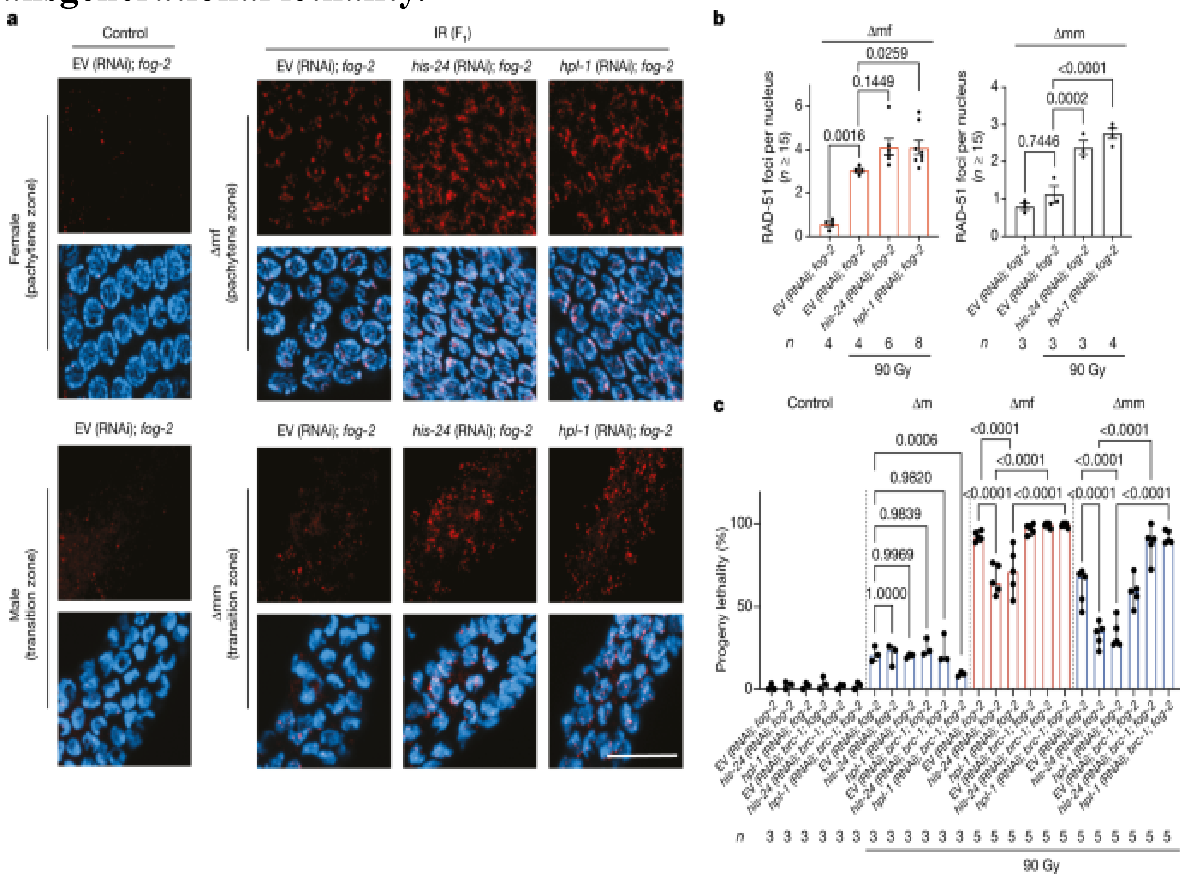
HPL-1 is essential for the establishment and maintenance of heterochromatin structure. Similar to *his-24* (RNAi); *fog-2*, HPL-1 knocked-down in *fog-2* mutants resulted in the loss of H3K9me2 in the germline of the F_1 worms (Fig. 4e). Consistently, we detected significant suppression of the progeny lethality of paternally irradiated *hpl-1; fog-2* mutant worms (Fig. 4f). We verified the restoration of viability by *hpl-1* RNAi in *fog-2* mutants (Extended Data Fig. 8c). Our observations suggest that the paternal DNA damage causes increased histone H1 expression, leading to increased heterochromatin formation, which in turn results in embryonic lethality in the F_2 generation.

HRR prevents transgenerational lethality

Heterochromatization could prevent faithful DSB repair, such as HRR[38,39]. There was a modest increase in RAD-51 foci, which indicate HRR[40], in the germline of F_1 worms compared with offspring from non-irradiated worms. Knockdown of *his-24* and *hpl-1* in *fog-2* mutants resulted in significantly increased RAD-51 foci in the germline of Δmm worms, whereas the

germline of Δmf worms showed a significant, albeit mild, increase in RAD-51 foci in *hpl-1* (RNAi); *fog-2* mutants (Fig. 5a,b). The increase in RAD-51 foci was most obvious in the pachytene zone of the Δmf germline and in the transition zone of the Δmm germline.

Fig. 5: Depletion of histone H1 or HPL-1 triggers HRR to alleviate transgenerational lethality.



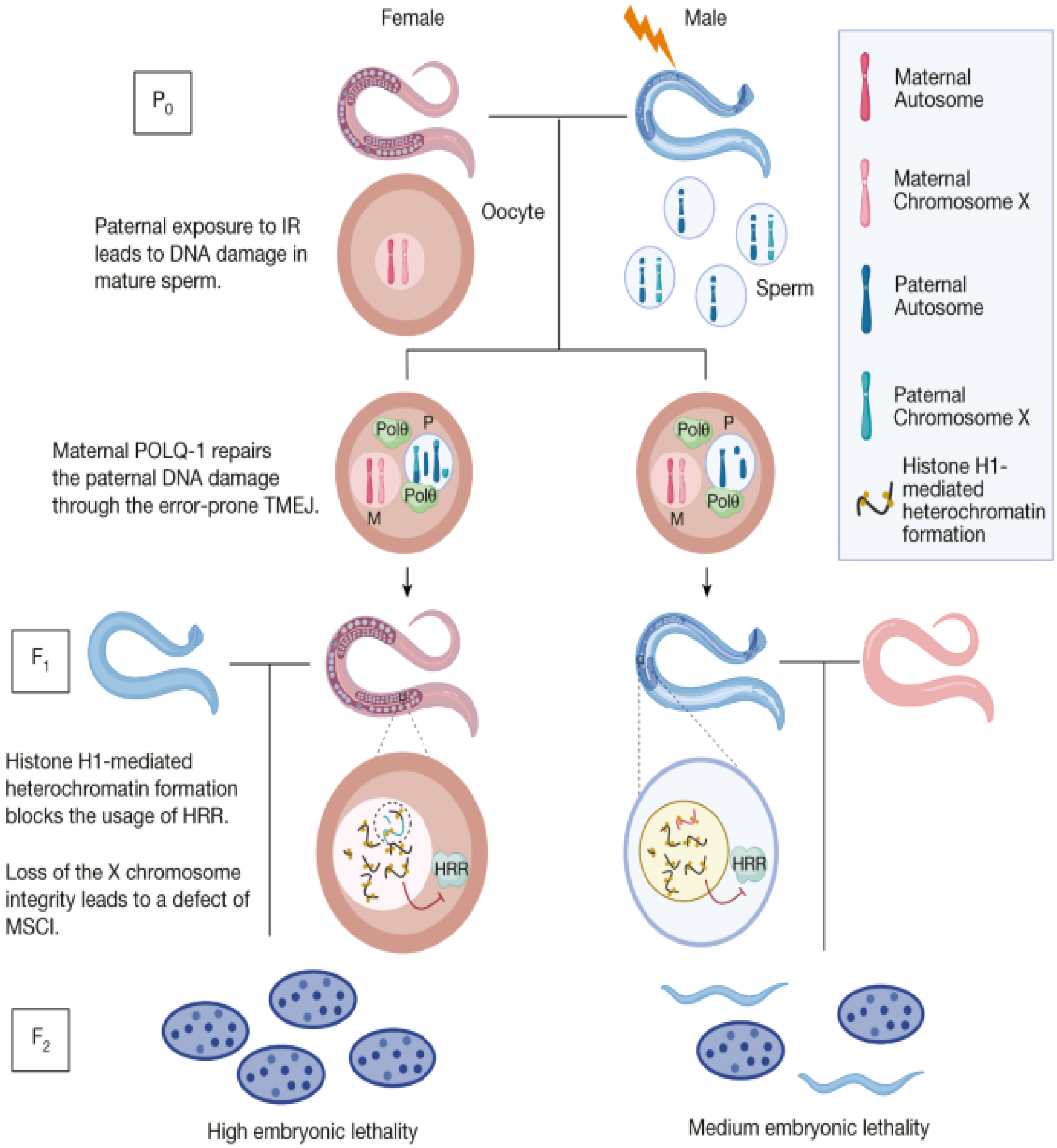
a, RAD-51 immunofluorescence (red) and DAPI staining (blue) in the germlines of EV (RNAi); *fog-2*, *his-24* (RNAi); *fog-2* and *hpl-1* (RNAi); *fog-2* worms with or without paternal exposure to 90 Gy ionizing radiation. Top, pachytene zone of the female germlines. Bottom, transition zone of male germlines. Scale bar, 10 µm. Experiment in **a** was repeated three times with similar observations. **b**, Quantification of RAD-51 foci per nucleus in the germline of females and males, with or without paternal exposure to 90 Gy ionizing radiation. More than 15 nuclei were counted for each germline; *n* represents the number of germlines. Each dot shows the mean of the number of RAD-51 foci per nucleus in one germline. Bars show median ± 95% confidence interval. *P* values are shown; nested one-way ANOVA with

Bonferroni's multiple comparisons test, reflecting the variation between different germlines. **c**, Progeny lethality characterization of empty vector, *his-24* and *hpl-1* RNAi on *fog-2* and *brc-1; fog-2* mutants with or without paternal ionizing radiation irradiation (90 Gy). Control, Δm: $n = 3$; Δmf, Δmm: $n = 5$ biological replicates. Data are median ± 95%confidence interval. P values are shown; GLM with logit link function and Tukey multiple comparisons; one-way ANOVA with arcsine transformation was used for the proportional data to confirm the statistical results and are included in the full statistical results (Supplementary Table 1).

To assess whether the rescue effect in the worms devoid of HIS-24 and HPL-1 is owing to the activation of HRR, we knocked down *his-24* and *hpl-1* in the *brc-1; fog-2* double mutant. The rescue of progeny lethality upon *his-24* and *hpl-1* depletion by RNAi was completely abolished in *fog-2; brc-1* mutants, indicating that HRR mediates the restored viability of the F_1 progeny upon H1 depletion (Fig. 5c). We confirmed the requirement of HRR for the ability of HIS-24 and HPL-1 depletion to restore viability in hermaphrodite wild-type and *brc-1* mutants (Extended Data Fig. 8d). Although the depletion of BRC-1 completely abolished the progeny viability of the H1-depleted F_1 offspring, it did not affect the progeny viability of the P_0 generation, indicating that the H1-mediated DNA repair restriction mainly affects the F_1 germline.

We conclude that the paternally inherited DNA damage leads to increased histone H1 and heterochromatin formation, which in turn prevents the F_1 generation from using the faithful DSB repair machinery. Removal of the heterochromatin barrier facilitates the engagement of the error-free HRR and significantly rescues the transgenerational embryonic lethality caused by paternal ionizing radiation irradiation (Fig. 6).

Fig. 6: Schematic model for the transgenerational effect of the paternal exposure to ionizing radiation.



Ionizing radiation (IR) exposure of male *C. elegans* leads to DNA DSBs in the mature sperm, which carries the fragmented DNA into the fertilized (un-irradiated) oocyte. The maternal oocyte provides the TMEJ machinery to repair the paternal DNA damage, resulting in various chromosomal aberrations. In the germline of the F_1 offspring, linker histone H1-mediated heterochromatin structures accumulate in the chromatin of the germ cells, which prevents the DNA DSBs from being detected and repaired by the error-free HRR. Specifically in the female F_1 (Δmf), X chromosome

translocations and fragmentations prevent MSCI, thus further aggravating the embryonic lethality in the progeny. Together, paternal exposure to ionizing radiation leads to transgenerational embryonic lethality via heterochromatization-restricted DNA repair access.

Discussion

The determination of the transgenerational consequences of parental exposure to ionizing radiation has been limited mainly because the mechanisms underlying this process have remained unknown. Studies in animals have provided evidence that paternal exposure to ionizing radiation could result in compromised viability, fertility and genome stability in the offspring[41]. Bovine oocytes that are fertilized with ionizing radiation-treated mature sperm display chromosomal fragmentation and translocations of the paternal chromosomes during early embryogenesis, resulting in aneuploidy and unequal chromosome segregation[42]. Such genetic mosaicism is the leading cause of the high miscarriage rate and low implantation rate in in vitro fertilization[43,44]. Viable F_1 offspring of irradiated male mice display increased genomic instability and increased incidence of cancers[45,46]. The specific vulnerability of mature sperm is consistent with the correlation between the increased risk of malignant disease in children and paternal radiation exposure around the time of conception—that is, when mature sperm has been exposed to radiation[14,16]. Consistently, germline hypermutation in humans has recently been associated with paternal exposure to chemotherapeutic agents before conception[21], which might increase the risk of genetic disease in the offspring[22].

The specific vulnerability of mature sperm is explained by its highly compacted chromatin structure precluding DSB repair, necessitating repair by maternal factors after fertilization. Depletion of the HRR and NHEJ factors in mice could significantly increase chromosomal structural aberrations in early embryos carrying paternal DSBs[23]. We found that in *C. elegans*, maternal TMEJ is the major DSB repair mechanism acting on the paternal genome upon fertilization. The de novo germline structural variants generated under physiological conditions also show the TMEJ signature and in humans is particularly enriched in paternally generated DNMs[5,47,48].

From an evolutionary perspective, unlike the stable maternal genome, the unstable paternal genome contributes greatly to the genetic diversity of a species. TMEJ appears to have a major role in shaping genome evolution, as its signature is present both in *C. elegans* variants[32] and in paternal human structural variants. It is thus tempting to speculate that the mechanisms governing the engagement of maternal TMEJ in the paternal genome thereby contribute to genetic diversification.

Our observation that the genome instability observed in the progeny of the irradiated males is driven by changes in the epigenome provides a rationale for developing potential therapeutic approaches to prevent the detrimental heritable consequences of parental radiation exposure. Heterochromatin inhibitors or H3K9 methyltransferase inhibitors[49] might potentially be considered for promoting HRR in the progeny carrying paternal genome instability. In addition, our work further emphasizes that the paternal genome in mature sperm is particularly vulnerable, potentially giving rise to genome instability in the progeny, increasing the need for protection from mutagenic exposure particularly in the two months period pre-conception. In sum, our work elucidates a previously unknown mechanism underlying the transgenerational effect of paternal exposure to radiation and mandates further evaluation of the transgenerational effects of paternal DNA damage in humans.

Methods

C. elegans strains

All strains were maintained based on standard conditions at 20 °C. Strains used were N2 (Bristol; wild type), CB4108 *fog-2(q71)*, BC784 *spe-8(hc50)*, RB1067 *his-24(ok1024)*, MT13971 *hpl-1(n4317)*, DW102 *brc-1(tm1145)*, FX1524 *cku-70(tm1524)*, FX2026 *polq-1(tm2026)*, BJS1017 *his-24(ok1024); fog-2(q71)*, BJS1018 *hpl-1(n4317); fog-2(q71)*, BJS1019 *brc-1(tm1145); fog-2(q71)*, BJS1020 *cku-70(tm1524); fog-2(q71)*, BJS1021 *polq-1(tm2026); fog-2(q71)*.

Measurement of ionizing radiation-induced progeny lethality

For the feminized mutants, synchronized L4 females and males were separated and maintained overnight. On the second day, the adult females or males either remained untreated or were irradiated with the indicated dose of ionizing radiation inflicted by a caesium 137 irradiation source (Biobeam GM 8000, Eckert & Ziegler, Gamma-Service Medical). Afterwards, ≥3 irradiated adult worms and ≥3 non-irradiated opposite-sex adults were immediately transferred to 3 new plates served as 3 biological replicates and allowed to lay eggs for 2 h. Then the males were removed and left the females to continue egg-laying for another 4 h. The females were then removed and the number of eggs was counted. The number of surviving progeny was characterized 24 h later to examine the progeny lethality of the P_0 generation. Then, we transferred ≥3 surviving male progeny (F_1) or ≥3 surviving female progeny (F_1) to new plates, serving as one biological replicate, and placed them with the three untreated opposite-sex adults and allowed them to lay eggs for one day. At least three biological replicates were included in each experiment. The adults were then removed and the number of laid eggs was counted. Twenty-four hours later, the surviving progenies were counted to characterize the progeny lethality of the F_1 generation.

For the hermaphrodite worms, synchronized hermaphrodite late L4 were separated from the rest of the worms by picking, and ≥3 late L4 hermaphrodites were either untreated or irradiated with the indicated dose of ionizing radiation. Three irradiated hermaphrodites were transferred to 3 separate plates as 3 biological replicates and allowed to lay eggs for 6 h. The adults were then removed and the number of eggs was counted 24 h later, the hatched progeny were quantified as the progeny lethality of the P_0 generation. The surviving worms were transferred to three plates and allowed to lay eggs for one day. The adults were removed and the progeny lethality of the consequent generations was quantified 24 h later.

Developmental assay

Synchronized L1 worms of control and paternally treated F_1 were generated via standard hypochlorite treatment. Arrested L1 worms were placed on NGM agar plates with OP50, fed with bacteria and incubated at 20 °C for 48 h. Then the larval stage of worms was characterized under a Zeiss

discovery.V8 microscope. For each experiment, >30 L1 larvae were included for each replicate, $n = 3$ biological replicates were used.

Quantification of germ cell apoptotic corpses

Day-1 adult female worms were immobilized using 5 mM levamisole (AppliChem A431005) and mounted on a 2% agarose pad on a microscope slide. The number of apoptotic corpses was scored via Nomarski DIC microscopy on a Zeiss Axio Imager M1/2 based on the refractive morphological changes occurring in apoptotic germ cells within the gonad loop[50].

RNAi treatment

RNAi feeding clones were obtained from the library of J. Ahringer. The *E. coli* feeding strain HT115 (DE3) with RNAi clones were cultured with LB medium containing ampicillin (100 μg ml^{-1}) overnight. IPTG (1 mM) was added to the culture for the induction of RNAi product before seeded on RNAi agar plates (NGM agar with ampicillin and IPTG). For the RNAi feeding assay, >30 synchronized L1 larvae as P_0 generation were placed on the RNAi agar plates seeded with *E. coli* feeding strain HT115 (DE3) containing specific RNAi or empty vector control. Three days later, adult males and females were separated and transferred to fresh RNAi plates for maintaining the RNAi efficiency until further experiments were performed. The subsequent experiments were performed as described in 'Measurement of ionizing radiation-induced progeny lethality'.

Immunofluorescence staining

Adult worms were picked from plates and transferred to a drop of M9 buffer onto a 0.3% polylysine-treated three-well slide (3 × 14 mm printed wells slides from Fisher Scientific). Germline dissection was carried out with two syringe needles, followed by fixation with 3.7% formaldehyde for 1 h. Then, a 24 × 24 mm coverslip was placed onto the drop, and the slide was left in a −80 °C freezer for 10 min to perform the freeze-cracking procedure. Then the slide was quickly transferred to −20 °C methanol for less than 1 min. For visualizing the nuclei, slides were washed once with PBS and once with

PBST (0.2% Tween in PBS) and mounted with DAPI Fluoromount-G mounting medium (Southern Biotech) and sealed with nail polish. For the other staining, after fixation, slides were washed 1 time with 1× PBS and 2 times with 1× PBT (0.5% Triton X-100 in PBS). To improve the signal quality, slides were first blocked for 20 min with Image-iT FX signal enhancer (Thermo Fisher) before blocking with 1× PBT containing 10% donkey serum for another 20 min. Afterwards, primary antibodies diluted with 1× PBT containing 5% donkey serum were applied to the slides and incubated at 4 °C overnight. After washing 3 times with 1× PBT, the slides were incubated with secondary antibodies diluted with 1× PBT at 37 °C for 30 min. Then slides were washed with 1× PBT 3 times, mounted with DAPI Fluoromount-G mounting medium (Southern Biotech) and sealed with nail polish. Slides were stored at 4 °C in the dark before imaging.

Primary antibodies used for immunofluorescence staining were rabbit polyclonal anti-phospho-RNAPII (Ser2) antibody (Thermo Fisher, A300-654A; dilution 1:500 in PBT); mouse monoclonal anti-H3K9me2 antibody (Abcam, ab1220; dilution 1:100 in PBT); rabbit polyclonal anti-HIM-8 (Novus Biologicals, 41980002; dilution 1:100 in PBT); rabbit anti-RAD-51 antibody (a gift from the laboratory of A. Gartner; dilution 1:2,000 in PBT). Secondary antibodies used were AlexaFluor 488 donkey anti-mouse IgG (Thermo Fisher, A21202; dilution 1:500 in PBT) and AlexaFluor 594 donkey anti-rabbit IgG (Thermo Fisher A21207; dilution 1:500 in PBT).

Fluorescence images for quantification were taken with a Zeiss Meta 710 confocal laser scanning microscope. For quantification, fixed exposure time was set for different treatments and strains. For H3K9me2, RNAPII p-Ser2 and RAD-51 staining, *z*-stack images were acquired with Zeiss Meta 710 confocal microscope, and the H3K9me2 and RNAPII p-Ser2 signal intensity and the foci number of RAD-51 foci per nucleus were quantified with Imaris x64 9.1.2 software. Fluorescence intensities were normalized to DAPI signal.

SILAC assay

The stable isotope labelling procedure was described in a previous study[51]. In brief, ET505 *E. coli* (lysine auxotrophy, from Coli Genetic Stock Center)

were grown in M9 minimal medium (Na_2HPO_4 5.8 g l^{-1}, KH_2PO_4 3 g l^{-1}, NaCl 0.5 g l^{-1}, NH_4Cl_2 1 g l^{-1}, glucose 0.2% (w/v), $MgSO_4$ 1 mM, thiamine 0.01% (w/v) and 40 µg ml^{-1} $^{13}C_6$-labelled lysine (Cambridge isotope laboratory) or 40 µg ml^{-1} normal l-lysine, and incubated at 37 °C overnight to reach A_{600} = 1. Bacteria were concentrated to A_{600} = 50 and seeded to NGM-N plates (3 g of NaCl and 12 g of agarose dissolved in 970 ml deionised water).

Synchronized embryos generated by hypochlorite treatment were hatched and arrested in M9 buffer, and then L1 worms were transferred to NGM-N plates seeded with heavy isotope labelled lysine (heavy lysine)- or normal lysine (light lysine)-labelled ET505 *E. coli*. Worms were fed with labelled bacteria for two generations to reach the incorporation rate >97%, and then picked the late L4 stage worms to irradiate with ionizing radiation or mock-treated. Four replicates were included in this experiment, as two of them were the ionizing radiation-treated heavy lysine group and mock-treated light lysine group, whereas the other two replicates were the mock-treated heavy lysine group and ionizing radiation-treated light lysine group.

Equal numbers of the F_1 adult worms were washed off from heavy lysine plates and light lysine plates with M9 buffer and combined. After removing the M9 buffer, lysis buffer was added to the worm pellet (6 M guanidinium chloride (GuCl), 10 mM TCEP, 40 mM CAA, 100 mM Tris-HCl). Heat the sample at 95 °C for 10 min and sonicate the sample with Bioruptor (30 s sonication, 30 s break, 10 cycles, high performance). Heating and sonication were repeated once more. Then samples were centrifuged at 20,000*g* for 20 min, and the supernatant was collected. Five microlitres of protein solution was diluted with 20 mM Tris to reduce the concentration of guanidinium chloride to below 0.6 M. 50 mM TEAB was added to dilute the samples to 100 µl and then 1 µg Lys-C was added for incubation at 37 °C for 4 h. Samples were further diluted with 180 µl TEAB and treated with 2 µg Lys-C at 37 °C overnight. Enzyme digestion was stopped by adding formic acid to 1%, and the sample purification by StageTip was carried out according to CECAD/CMMC Proteomics Core facility's standard protocol (https://www.proteomics-cologne.com/protocols). The mass spectrometry

proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD031873.

Single-worm WGS

Single L4 male and female *fog-2*, and hermaphrodite wild-type with indicated treatment were transferred to plates with UV-killed OP50 bacteria, in order to reduce the contamination of bacterial DNA. On the second day, a single adult worm was picked to 10 µl of M9 buffer, then the samples were frozen at −80 °C. DNA was extracted following the standard Illumina DNA preparation protocol. Libraries were prepared for sequencing using the standard Illumina protocols. In brief, 120 ng of genomic DNA was tagmented with adaptor sequence using bead-linked transposomes. Tagmented DNA was amplified by PCR for 5 cycles. Libraries were sequenced on the Illumina HiSeq 2500 following the manufacturer's protocols. The data have been deposited with links to BioProject accession number PRJNA826255 in the BioProject database.

Telomere FISH

Telomere FISH was carried out by a modified protocol[52]. Twenty adult females or males were transferred to a drop of M9 buffer onto a 0.3% polylysine-treated 3-well slide (3 × 14 mm printed well slides; Fisher Scientific). Germline dissection was carried out with two syringe needles, followed by fixation with 3.7% formaldehyde for 1 h. Then a 24 × 24 mm coverslip was placed onto the drop, and the slide was left in a −80 °C freezer for 10 min to perform the freeze-cracking procedure. Then the slide was quickly transferred to −20 °C methanol for less than 1 min. The slides were washed once with 1× PBS and incubated in permeabilization buffer (0.5% Triton X-100 in 1× PBS) for 1 h at room temperature followed by a wash in 1× PBS. Then slides were quickly washed with 0.01 N HCL followed by a wash with 0.1 N HCL for 2 min. To prevent unspecific binding of the FISH probe, 50 µg ml^{-1} RNase A solution (10 µg ml^{-1} RNase A in 1× PBS) was added to the slide and incubated at 37 °C for 45 min. Afterwards, slides were washed 2 times with 2× SSC. For pre-hybridization, 50 µl of pre-hybridization solution (2× SSC, 50% formamide) was added on the slides and incubated in a humid chamber for 2 h at room temperature. Then the

FISH probe (PNA-FISH TTAGGC telomeric probe, Panagene, resuspended to 100 µM, fluorophore: Cy3) was diluted as 1:500 in hybridization buffer (2× SSC, 50% formamide, 10% (w/v) dextran sulfate, 50 µg ml^{-1} heparin, 100 µg ml^{-1} sheared salmon sperm DNA). After pre-hybridization, the solution was removed from the slide as much as possible, then 50 µl of FISH probes were added to the sample and covered with Frame-Seal in situ PCR and Hybridization Slide Chamber (Bio-Rad SLF0601), then the slides were incubated over-night at 37 °C. On the second day, the samples were denatured for 5 min at 80 °C and continued incubating at 37 °C for 2 days. Afterwards, the hybridization chambers were removed, and slides were washed in 2× SSC for 5 min at room temperature. To fixate the staining, the slides were washed 3 times in preheated 2× SSC at 37 °C, and twice in preheated 0.2× SSC at 55 °C. As the last step, slides were washed in 1× PBT for 10 min at room temperature and mounted by 7 µl Vectashield mounting medium containing DAPI (Vector Laboratories H-1200-10).

Staining images were taken with a Zeiss Meta 710 confocal laser scanning microscope was used. To visualize the telomeric signal, *z*-stack images were acquired with a Zeiss Meta 710 confocal microscope, and the *z*-stack pictures were processed and the number of DNA fragments was counted with Image J/Fiji v2.3.0/1.53f.

Public datasets

The following public datasets have been re-analysed: The deletions of the hard-filtered variants of the 20220216 CeNDR 29 release[30] were downloaded from https://www.elegansvariation.org/data/release/20220216. The data for the *C. elegans* mutation accumulation experiment were downloaded from the supplementary data from Volkova et al.[31]. The filtered hg38 SNV_INDEL_SV_phased_panel.vcf files for all chromosomes from the 20220422 release of the 1000 Genomes Project[34] were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/. The hg38 illumina-polaris-v2.1-sv-truthset structural variants (https://github.com/Illumina/Polaris) were downloaded from https://s3-us-west-1.amazonaws.com/illumina-polaris-v2.1-sv-truthset/all_merge.vcf.gz. The processed hg38 variants of the 1,548 trios from Iceland including

gamete-of-origin analysis were downloaded from the supplementary data from Jónsson et al.[35].

Bioinformatics analysis

Gene set enrichment analysis

The enrichment analysis for chromosomal gene distributions was done in R v3.6.3 with the GSEA function of clusterProfiler v3.14.3[53] was used with maxGSSize = 20000 and nPerm = 20000.

WGS preprocessing

The fastq files were preprocessed with Fastp v0.20.0[54], and mapped with BWA-0.7.17[55] with the parameters bwa mem -M -K 100000000, and the reference genome ce11. The mapped files were converted to BAM and sorted with samtools v1.6[56], and duplicated reads were removed with GATK v4.1.0.0 MarkDuplicates[57].

Structural variant calling

Structural variants were called with Manta v1.6.0[58] and only structural variants that passed all of the Manta quality filters were used. To find structural variants that are newly induced in the F_1 generation, structural variants that overlapped with any structural variant of any P_0 sample were filtered out (full structural variation sites with or without filtering in Supplementary Table 3). Repeat regions are difficult to map and identify, we therefore deleted any insertion–deletion mutant within a repeat region, or any translocation for which both break points overlapped with the same repeat class. Manta calls translocations in both directions as two break points including a position confidence interval. To avoid duplicates, we filtered translocations that had the same start and end break point within the respective confidence interval in any combination.

Translocation types

A translocation is called as two break points and can appear in four different ways in a VCF file. The reference sequence s is replaced by the sequence t after the fusion to position p, respective before the fusion at position p. This can happen in four ways:

- Type 1: t[p[The genomic location extending right from the position p is fused after t. In other words, these are fusions between the 3′ sense strand with the 5′ sense strand.
- Type 2: t]p] The reverse component of the genomic location left of the position p is fused after t. In other words, these are fusions between the 3′ sense strand and the 5′ anti-sense strand.
- Type 3:]p]t The genomic location extending left from the position p is fused before t. This is the same as Type 1.
- Type 4: [p[t The reverse component of the genomic location extending right from the position p is fused before t. In other words, these are fusions between the 3′ anti-sense strand and the 5′ sense strand.

See https://github.com/samtools/hts-specs/blob/master/VCFv4.1.pdf for further details.

Circos plots

The library circlize v0.4.12[59] in R v3.6.3 was used to generate circos plots.

Templated insertions with distribution

The inserted sequence between the break points was searched within ±25 bp around the break points in both directions, in the normal orientation, as well as in the reverse, complement, and reverse complement orientation. Insertions ≥3 bp for which a template could be found within ±25 bp were called templated insertions, while any other insertion was classed as miscellaneous.

Microhomology with permutations

For each translocation 8 bp surrounding both break sites (that is, 4 bp upstream and 4 bp downstream of both break sites) were compared in an 8 × 8 grid (that is, each of the surrounding bases is compared to every other base). Matching bases were scored 1 and nonmatching bases were scored 0. One map therefore contains a 1 for each of the 64 combinations that have the same base, and 0 otherwise. The heat maps shown in the figure contain the sum of all such respective heatmaps divided by the total number of translocations. For each of the four translocation classes a separate microhomology was calculated. For type 1 translocations the left and right flank are both 5′ to 3′ on the sense strand. The left flank of type 2 translocations is the 5′ to 3′ sense strand, while the right flank is the reverse complement sequence. For type 4 translocations the left flank is the reverse complement sequence, while the right flank is the 5′ to 3′ sense strand sequence.

To calculate the significance of individual bins a permutation test was done. For each permutation the same number of translocations (of the same type) as in the original heatmap was randomly distributed on the genome to calculate the microhomology ratios for each of the 64 bins. To calculate a *P* value a permutation test was calculated with 100,000 permutations. To calculate the adjusted *P* value for the 64 bins statsmodels v0.11.1[60] multipletests methods with the parameter method='fdr_bh' in Python 3.6[61] was used.

Base composition

For all translocations with a microhomology of length 1 ($n = 35$ for *fog-2*, and $n = 58$ for wild type) the base composition for 8 bp around the break points was calculated. For each of the 8 positions the percentage of A, C, T and G was calculated. To be able to compare it to a random background distribution we sampled the same number of positions—that is, 35 for *fog-2* and 58 for wild type—and calculated the average percentage of each A, C, T, and G for 25.000 such permutations.

Analyses of public *C. elegans* datasets

CeNDR

The deletions of the hard-filtered variants of the 20220216 CeNDR[30] release were downloaded and further filtered for deletions between 8 and 200 bp. Deletions for which both break sites were annotated within the same repeat class were removed. Each deletion got categorized into non-homology, that is, no matching base at the break sites, microhomology—that is, exactly one matching base at the break sites, and macro-homology—that is, more than one matching base. By chance we would expect 75% of break sites to be non-homologous, 16.66% microhomologous, and 8.33% macro-homologous. The over-representation of microhomologous deletions sites were calculated with the binomial test function binom_test in Python's Scipy-v1.5.1 package. The over-represented microhomologous variants were used in the heat map as described above for a 16 × 16 grid.

Mutation accumulation

The data for the mutation accumulation experiment were downloaded from the supplementary data from Volkova et al.[31]. Deletions for which both break sites were annotated within the same repeat class were removed, and only deletions with a length between 8 and 200 bp were considered for the subsequent analysis. The statistics and heat map were calculated as described above.

Analyses of public human datasets

1000 Genomes Project

The filtered hg38 SNV_INDEL_SV_phased_panel.vcf files for all chromosomes from the 20220422 release of the 1000 Genomes Project[34] were downloaded and further filtered for deletions between 8 and 200 bp. Since the microhomology footprint of humans is 2–6 bp, we defined the 3 categories different from *C. elegans*. 0–1 bp homology is expected by chance in 75% + 16.66% = 91.66% of break sites. Microhomology—that is, 2–6 bp homology, in 8.325% of break sites, and macrohomology in 1/12288 ≈ 0.00008% of break sites. The statistics and heat map were otherwise calculated as described above.

Polaris

The hg38 illumina-polaris-v2.1-sv-truthset structural variants (https://github.com/Illumina/Polaris) were downloaded and filtered for deletions that had a PASS in the quality column. To focus on de novo deletions, any deletion that overlapped with either parent got filtered out, and only deletions between 8 and 200 bp were considered for the subsequent analysis. The statistics and heat map were calculated as described above.

Iceland trios

The processed hg38 variants of the 1548 trios from Iceland including gamete-of-origin analysis were downloaded from the supplementary data from Jónsson et al.[35]. Only deletions between 8 and 200 bp for which the gamete of origin was available were considered. The statistics and heat map were calculated as described above separately for deletions coming from the mother and father.

Data presentation and statistical analysis

All the data and statistical significances were analysed using the GraphPad Prism 7 software package (GraphPad) and R studio. For the proportion data shown in this paper, GLM with logit link function (R v4.0.2 and emmeans v1.5.2 (https://cran.r-project.org/web/packages/emmeans/index.html)) and ordinary ANOVA with arcsine transformed value (arcsine transformation equation: $Y = \arcsin(\sqrt{(Y/n)}) \times 180/\pi$) were both applied to confirm the significance of the observations, and the full statistic results are shown in the Supplementary Table 1. In addition, a QQ plot was attached for the ANOVA analysis to assess the normal distribution of the transformed value. Only the *P* values calculated from the GLM method are shown in the figures. Statistical methods, *P* values, sample size information and error bar descriptions are reported in the figure legends. Randomization was not applied because the group allocation was guided based on the genotype of the respective mutant worms. Worms of a given genotype were nevertheless randomly selected from large strain populations for each experiment without any preconditioning. Blinding was not applied as the experiments were carried out under highly standardized and predefined conditions such that an

investigator-induced bias can be excluded. For progeny lethality characterization and staining quantifications, median with 95% confidence interval was used as these data types contain outliers.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The single-worm WGS data have been deposited in the BioProject database under accession number PRJNA826255. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD031873.

Code availability

Custom code used in this study is available from https://github.com/Meyer-DH/swWGS_Custom_Code.

References

1. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).

2. UNSCEAR. *Hereditary Effects of Radiation. 2001 Report to the General Assembly, with Scientific Annex* (United Nations, 2001).

3. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the ageing process. *Nature* **592**, 695–703 (2021).

4. Vijg, J. & Dong, X. Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell* **182**, 12–23 (2020).

5. Wright, S. The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution. In: *Proc. of the XI International Congress of Genetics* **8**, 209–222 (1932).

6. Hu, C. et al. Association between inherited germline mutations in cancer predisposition genes and risk of pancreatic cancer. *JAMA* **319**, 2401 (2018).

7. Kloosterman, W. P. et al. Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015).

8. Belyeu, J. R. et al. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am. J. Hum. Genet.* **108**, 597–607 (2021).

9. Malaspina, D. Paternal factors and schizophrenia risk: de novo mutations and imprinting. *Schizophr. Bull.* **27**, 379–393 (2001).

10. Cawthon, R. M. et al. Germline mutation rates in young adults predict longevity and reproductive lifespan. *Sci. Rep.* **10**, 10001 (2020).

11. Holtgrewe, M. et al. Multisite de novo mutations in human offspring after paternal exposure to ionizing radiation. *Sci. Rep.* **8**, 14611 (2018).

12. Dubrova, Yu. E. Mutation induction in the mouse and human germline. *Russ. J. Genet.* **52**, 17–28 (2016).

13. Little, M. P., Goodhead, D. T., Bridges, B. A. & Bouffler, S. D. Evidence relevant to untargeted and transgenerational effects in the offspring of irradiated parents. *Mutat. Res.* **753**, 50–67 (2013).

14. Gardner, M. J. et al. Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria. *Brit. Med. J.* **300**, 423–429 (1990).

15. Wakeford, R. Childhood leukaemia and radiation exposure of fathers-the end of the road, perhaps? *J. Radiol. Prot.* **23**, 359–362 (2003).

16. Gardner, M. J., Hall, A. J., Downes, S. & Terrell, J. D. Follow up study of children born elsewhere but attending schools in Seascale, West Cumbria (schools cohort). *Brit. Med. J.* **295**, 819–822 (1987).

17. Yeager, M. et al. Lack of transgenerational effects of ionizing radiation exposure from the Chernobyl accident. *Science* **372**, 725–729 (2021).

18. Morton, L. M. et al. Radiation-related genomic profile of papillary thyroid carcinoma after the Chernobyl accident. *Science* **372**, eabg2538 (2021).

19. Kodaira, M., Satoh, C., Hiyama, K. & Toyama, K. Lack of effects of atomic bomb radiation on genetic instability of tandem-repetitive elements in human germ cells. *Am. J. Hum. Genet.* **57**, 1275 (1995).

20. Kodaira, M. et al. No evidence of increased mutation rates at microsatellite loci in offspring of A-bomb survivors. *Radiat. Res.* **173**, 205–213 (2010).

21. Kaplanis, J. et al. Genetic and chemotherapeutic influences on germline hypermutation. *Nature* **605**, 503–508 (2022).

22. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).

23. Marchetti, F., Essers, J., Kanaar, R. & Wyrobek, A. J. Disruption of maternal DNA repair increases sperm-derived chromosomal aberrations. *Proc. Natl Acad. Sci. USA* **104**, 17725–17729 (2007).

24. McClintock, B. The stability of broken ends of chromosomes in *Zea mays*. *Genetics* **26**, 234–282 (1941).

25. Gartner, A., Milstein, S., Ahmed, S., Hodgkin, J. & Hengartner, M. O. A conserved checkpoint pathway mediates DNA damage-induced apoptosis and cell cycle arrest in *C. elegans*. *Mol. Cell* **5**, 435–443 (2000).

26. Rappaport, Y. et al. Bisection of the X chromosome disrupts the initiation of chromosome silencing during meiosis in *Caenorhabditis*

elegans. *Nat. Commun.* **12**, 4802 (2021).

27. Wyatt, D. W. et al. Essential roles for polymerase θ-mediated end joining in the repair of chromosome breaks. *Mol. Cell* **63**, 662–673 (2016).

28. Roerink, S. F., van Schendel, R. & Tijsterman, M. Polymerase theta-mediated end joining of replication-associated DNA breaks in *C. elegans*. *Genome Res.* **24**, 954–962 (2014).

29. Burma, S., Chen, B. P. C. & Chen, D. J. Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. *DNA Repair* **5**, 1042–1048 (2006).

30. Cook, D. E., Zdraljevic, S., Roberts, J. P. & Andersen, E. C. CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic Acids Res.* **45**, D650–D657 (2017).

31. Volkova, N. V. et al. Mutational signatures are jointly shaped by DNA damage and repair. *Nat. Commun.* **11**, 2169 (2020).

32. Van Schendel, R., Roerink, S. F., Portegijs, V., Van Den Heuvel, S. & Tijsterman, M. Polymerase Θ is a key driver of genome evolution and of CRISPR/Cas9-mediated mutagenesis. *Nat. Commun.* **6**, 7394 (2015).

33. Carvajal-Garcia, J. et al. Mechanistic basis for microhomology identification and genome scarring by polymerase theta. *Proc. Natl Acad. Sci. USA* **117**, 8476–8485 (2020).

34. Byrska-Bishop, M. et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *SSRN Electron. J.* **185**, 3426–3440.e19 (2021).

35. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).

36. Jedrusik-Bode, M. Histone H1 and heterochromatin protein 1 (HP1) regulate specific gene expression and not global transcription. *Worm* **2**, e1002940 (2013).

37. Studencka, M. et al. Transcriptional repression of Hox genes by *C. elegans* HP1/HPL and H1/HIS-24. *PLoS Genet.* **8**, e1002940 (2012).

38. Lorković, Z. J. & Berger, F. Heterochromatin and DNA damage repair: use different histone variants and relax. *Nucleus* **8**, 583–588 (2017).

39. Cann, K. L. & Dellaire, G. Heterochromatin and the DNA damage response: the need to relax. *Biochem. Cell Biol.* **89**, 45–60 (2011).

40. Alpi, A., Pasierbek, P., Gartner, A. & Loidl, J. Genetic and cytological characterization of the recombination protein RAD-51 in *Caenorhabditis elegans*. *Chromosoma* **112**, 6–16 (2003).

41. Dubrova, Y. E. & Sarapultseva, E. I. Radiation-induced transgenerational effects in animals. *Int. J. Radiat. Biol.* **98**, 1047–1053 (2020).

42. Middelkamp, S. et al. Sperm DNA damage causes genomic instability in early embryonic development. *Sci. Adv.* **6**, eaaz7602 (2020).

43. Taylor, T. H. et al. The origin, mechanisms, incidence and clinical consequences of chromosomal mosaicism in humans. *Hum. Reprod. Update* **20**, 571–581 (2014).

44. Spinella, F. et al. Extent of chromosomal mosaicism influences the clinical outcome of in vitro fertilization treatments. *Fertil. Steril.* **109**, 77–83 (2018).

45. Lord, B. et al. Tumour induction by methyl-nitroso-urea following preconceptional paternal contamination with plutonium-239. *Br. J. Cancer* **78**, 301–311 (1998).

46. Vorobtsova, I. E., Aliyakparova, L. M. & Anisimov, V. N. Promotion of skin tumors by 12-*O*-tetradecanoylphorbol-13-acetate in two generations of descendants of male mice exposed to X-ray irradiation. *Mutat. Res. Mol. Mech. Mutagen.* **287**, 207–216 (1993).

47. Makova, K. D. & Li, W.-H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626 (2002).

48. Goldmann, J. M. et al. Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48**, 935–939 (2016).

49. Roberti, A., Valdes, A. F., Torrecillas, R., Fraga, M. F. & Fernandez, A. F. Epigenetics in cancer therapy and nanomedicine. *Clin. Epigenetics* **11**, 81 (2019).

50. Gartner, A., MacQueen, A. J. & Villeneuve, A. M. in *Checkpoint Controls and Cancer* vol. 280 (ed. Schönthal, A. H.) 257–274 (Humana, 2004).

51. Larance, M. et al. Stable-isotope labeling with amino acids in nematodes. *Nat. Methods* **8**, 849–851 (2011).

52. Seo, B. & Lee, J. Observation and quantification of telomere and repetitive sequences using fluorescence in situ hybridization (FISH) with PNA probes in Caenorhabditis elegans. *J. Vis. Exp.* **114**, e54224 (2016).

53. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

54. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

55. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).

56. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

57. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media, 2020).

58. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

59. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

60. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with Python. in *Proc. of the 9th Python in Science Conference* vol. 57 (SciPy, 2010).

61. vanRossum, G. *Python Reference Manual* (CWI, 1995).

Acknowledgements

We thank the CECAD proteomics and imaging facilities, and the Regional Computing Center of the University of Cologne (RRZK) for providing computing time and support on the DFG-funded High Performance Computing (HPC) system CHEOPS. Worm strains were provided by the National Bioresource Project (supported by The Ministry of Education, Culture, Sports, Science and Technology, Japan), the *Caenorhabditis* Genetics Center (funded by the NIH National Center for Research Resources), and the *C. elegans* Gene Knockout Project at the Oklahoma Medical Research Foundation (part of the International *C. elegans* Gene Knockout Consortium). NGS was performed at the West German Genome Center Cologne, supported by the DFG Research Infrastructure (407493903) of the NGS Competence Network (project 423957469). B.S. received the DFG grant SCHU 2494/7-1 as part of the DFG sequencing call. D.H.M. was supported by the Cologne Graduate School of Ageing Research. B.S. acknowledges funding from the Deutsche Forschungsgemeinschaft (SCHU 2494/3-1, SCHU 2494/7-1, SCHU 2494/10-1, SCHU 2494/11-1, SCHU 2494/15-1, CECAD EXC 2030 – 390661388, SFB 829, KFO 286, KFO 329 and GRK 2407), the Deutsche Krebshilfe (70114555), the H2020-MSCA-ITN-2018 (Healthage and ADDRESS ITNs) and the John Templeton Foundation Grant (61734). Schematics in the figures were created with BioRender.com.

Funding

Open access funding provided by Universität zu Köln.

Author information

Authors and Affiliations

1. Institute for Genome Stability in Aging and Disease, Medical Faculty, University Hospital and University of Cologne, Cologne, Germany

   Siyao Wang, David H. Meyer & Björn Schumacher

2. Cologne Excellence Cluster for Cellular Stress Responses in Aging-Associated Diseases (CECAD), Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany

   Siyao Wang, David H. Meyer & Björn Schumacher

Contributions

D.H.M. performed all bioinformatics analyses. S.W. performed and analysed the experiments and together with B.S. conceived and designed the study, and all authors wrote the manuscript.

Corresponding authors

Correspondence to Siyao Wang or Björn Schumacher.

Ethics declarations

Competing interests

B.S. is a co-founder of Agevio Therapeutics.

Peer review

Peer review information

Nature thanks Lenhard Rudolph and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 IR exposure of hermaphrodite WT in the late L4 stage leads to extended transgenerational embryonic lethality.

a. Transgenerational lethality characterization for IR treated WT hermaphrodites at the L3 (start of spermatogenesis), young L4 (ongoing spermatogenesis) and late L4 stage (mature sperm is stored in the spermatheca). Ten worms were included for each biological replicate, and n = 3 biological replicates. Error bars indicate median with 95% CI. **b**. Progeny lethality characterization for multiple generations upon IR treatment at the late L4 stage of hermaphrodites P0. Ten worms for each dosage group were included for each biological replicate, and n = 3 biological replicates. Error bars indicate median with 95% CI.

Extended Data Fig. 2 Paternal exposure to IR leads to genome instability in the F1 generation.

a. Quantification of "comet" like sperm (p-values and n number indicating sperm are shown, Chi-square test was performed). **b**. Quantification of the chromosomal aberrations in the F1 of paternal and maternal exposure to IR (p-values and n number indicating embryos are shown, Fisher's exact test was performed). **c**. Representative fluorescence microscopic images of DAPI (blue) in the *fog-2* embryos derived from the control or irradiated females. Zoomed image of chromosomal fragmentation is shown below, fragmented DNA is indicated with arrowheads. Scale bar = 10 µm. **d**.

Quantification of the intestinal chromosomal bridges in the F1 progenies of irradiated males and females (n *fog-2* = 9, n *spe-8* = 10 intestines, medians with 95% CI are shown, p-values are shown, two-factor GLM with logit link function and Tukey multiple comparisons were used). Two-way ANOVA with arcsine transformation was also used for the proportion data to confirm the statistical results and included in the full statistic results shown in Supplementary Table 1. **e**. Representative fluorescence microscopic images of hermaphrodite WTs' intestinal cells expressing H2B::GFP with or without parental IR irradiation (90 Gy). Chromosomal bridging was observed in the F1's intestinal cells. Scale bar = 10 μm. **c** and **e** were repeated three times with similar results. **f**. Developmental stages were determined 48h post L1 stage of control group of *fog-2* and *spe-8* and the F1 generation of *fog-2* and *spe-8* with paternal IR exposure, and n = 30 individuals were analyzed per experiment (n = 3 independent experiments per strain, p-values are shown, two-tailed t-test was performed). Full statistic results are shown in Supplementary Table 1. **g**. Quantification of apoptotic corpses per germline of unirradiated control, 90Gy treated Δf and Δmf of WT, *fog-2* and *spe-8* mutants (n = 9 germline, Medians with 95% CI are shown, p-values are shown, Two-way ANOVA with Bonferroni's multiple comparisons test were used). Full statistic results are shown in Supplementary Table 1. **h**. Number of new chromosomal translocations in Δmf (n = 6 worms) and Δmm (n = 8 worms). The box depicts the Median with the top and bottom quartiles, the whiskers extend to 1.5 IQR. New translocations and mapped reads per sample are shown in Supplementary Table 2. **i**. and **j**. Circos plots showing the novel chromosomal translocations in six *fog-2* Δmf (**i**) and eight *fog-2* Δmm (**j**). The list of all translocations is shown in Supplementary Table 3.

Extended Data Fig. 3 Gene set enrichment analysis on Chromosome I to Chromosome V of SILAC proteomic data from hermaphrodites WT (F1) with parental IR exposure.

Gene set enrichment analysis (GSEA) of SILAC proteomic data from hermaphrodites WT (F1) with parental IR exposure (90 Gy) for chromosomes I to V. The top of each panel shows the log2 fold changes of the proteomics data in a ranked order. The bottom panels show the running enrichment score as a green line. The dashed red line depicts the positions

furthest from 0. None of the chromosomes, except the X chromosome, is significantly different from a random distribution. The full statistics for each chromosome is provided in the Supplementary Table 4.

Extended Data Fig. 4 IR exposure in the late L4 stage of hermaphrodite WT leads to chromosomal rearrangement.

a. Whole-genome sequencing analysis shows the number of new chromosomal translocations in the hermaphrodites WT adults (F1) with parental IR exposure (90 Gy), normalized to untreated control. n = 11 worms, the box depicts the Median with the top and bottom quartiles, while the whiskers extend to 1.5 IQR. The number of new translocations and of mapped reads per sample is shown in Supplementary Table 2. **b**. Characterization of the DAPI-stained bodies in the late diakinesis oocytes of hermaphrodites WT adults (F1) with or without parental IR exposure (90 Gy). **c**. Merged images of immunofluorescence co-staining RNA Pol ll p-Ser2 (green), HIM-8 (red) and DAPI (blue) of dissected germlines from hermaphrodite WT adult (F1) with or without parental IR exposure (90 Gy). Three representative nuclei are shown on the right side. Scale bar = 20 µm. This experiment was repeated three times with similar results.

Extended Data Fig. 5 Circos plots showing the chromosomal translocations in ten hermaphrodites WT (F1) with parental IR (90 Gy) exposure.

The list of all translocations is shown in Supplementary Table 3.

Extended Data Fig. 6 Hermaphrodite WT animals use TMEJ to repair the inherited DNA damage.

a. Density plot showing the distribution of the template positions of the insertions found in between the translocation sites of *fog-2* adults (F1) with paternal exposure to IR (90 Gy). Only insertions are included that are at least 3 bp and whose templates were found within ±25 bp around one of the breakpoints (corresponding to the templated insertions in Fig. 3A). The orange ticks on the x-axis show the positions where the templates were

found. The blue curve depicts the kernel density estimation as a non-parametric way to estimate the probability density function of the found positions. The black lines indicate the distance cutoff of ±25 bp from the breakpoints. **b**. Distribution of translocation footprints in the WT adults (F1) with paternal exposure to IR (90 Gy). Templated inserts are insertions of ≥ 3 bp in between the fusion sites that have a matching sequence within ±25 bp around one of the 2 breakpoints. Miscellaneous insertions are insertions < 3 bp or insertions with no matching sequence within ±25 bp around the breakpoints. **c**. The same density plot as in (A) but for WT adults (F1) with paternal exposure to IR (90 Gy) (corresponding to the templated insertions in Extended Data Fig. 3B). **d**. Schematic illustration of the type 2 (top) and type 4 (bottom) translocations. The sense strand is in blue, while the anti-sense strand is in red. The numbers indicate the nucleotide distance to the breakpoints. Type 2 translocations are fusions between the 3′ sense strand with the 5′ anti-sense strand. Type 4 translocations are fusions between the 3′ anti-sense strand with the 5′ sense strand. The yellow boxes indicate the positions that show a microhomology. Type 2 shows a microhomology of the breakpoint of the left flank with the second base of the negative strand of the right flank. Type 4 shows a microhomology of the second to last base of the excluded part of the right flank, and the last base of the excluded part of the left flank. The microhomology of type 4 is therefore not in the regions that get fused. **e**. The heatmaps represent the sum of all type 2 (top, n = 51 translocations), respective type 4 (bottom, n = 45 translocations), translocation maps derived from the *fog-2* adults (F1) with paternal exposure to IR (90 Gy), excluding translocations with a templated insertion. Color-coded is the ratio of similarity. A darker color shows a higher sequence similarity between the corresponding bases. The numbers on the x- and y-axes are the same as in the explanation in Extended Data Fig. 3D. **f**. Similar to (e) but for WT adults (F1) with parental exposure to IF (90 G y). Type 1 (n = 65 translocations), type 2 (n = 57 translocations) and type 4 (n = 58 translocations) translocation maps derived from the WT adults (F1) with parental exposure to IR (90 Gy), excluding translocation with a templated insertion. **g**. Base composition around the breakpoints of the translocations with a microhomology of length 1 for *fog-2* (left, n = 35) or WT (right, n = 58). Position 0 indicates the break point. Negative numbers are bases not included after the fusion, while positive numbers depict the distance from the fusion site and are included after the fusion. Color-coded are the 4 bases,

the shadow in the background shows the distribution of AT, respective CG, for 25.000 permutations. The y-axis shows the percentage of the corresponding base at the given position. **h**. Knockdown of *EV, brc-1, cku-80* and *polq-1* by RNAi on the *fog-2* mutant with or without paternal IR exposure. n indicating biological replicates are shown. Medians with 95% CI are shown (p-values are shown. GLM with logit link function and Tukey multiple comparisons were used). One-way ANOVA with arcsine transformation was also used for the proportion data to confirm the statistical results and included in the full statistical results shown in Supplementary Table 1.

Extended Data Fig. 7 Analysis of public *C. elegans* and human datasets.

The heatmaps represent the sum of all base similarities of the microhomology deletion break sites. Color-coded is the ratio of similarity. A darker color shows a higher sequence similarity between the corresponding bases. The x-axis is the right, the y-axis the left flank of the deletion break site. Positive values indicate the flanking sequences, negative values the deleted sequences, i.e. -1 is the first deleted base **a**. The heatmap represents the sum of deletion sites with a length ≤ 200 bp and a 1 bp-microhomology of 540 natural isolates of the CeNDR dataset[30] (22.6% show a 1 bp-microhomology, expected are 16.66 %, binomial p-value = 5e-324, total deletions n = 68062). **b**. The heatmap shows the sum of deletion sites with a length ≤ 200 bp and a 1 bp-microhomology of samples without mutagen treatment of the public mutation accumulation dataset[31] (24.5% show a 1 bp-microhomology, 16.66 % expected, binomial p-value = 0.00035, total deletions n = 314). **c**. The heatmap represents the sum of deletion sites with a length ≤ 200 bp and a 2-6 bp microhomology of the human 1000 genome project[34] (28% show a 2-6 bp-microhomology, expected 8.3%, binomial p-value = 1e-323, total deletions n = 501024). **d**. The heatmap shows the sum of the *de novo* deletion sites with a length ≤ 200bp and a 2-6 bp microhomology of the *de novo* structural variants in the children of trios of the Polaris dataset (https://github.com/Illumina/Polaris) (19.5% show a 2-6 bp-microhomology, expected 8.3%, binomial p-value = 1.36e-292, total deletions n = 10883). **e**. The heatmap shows the sum of the *de novo* paternally-induced deletions with a length ≤ 200 bp and a 2-6 bp

microhomology of 1548 trios from Iceland[35] with known gamete-of-origin (44.2% show a 2-6 bp-microhomology, expected 8.3%, binomial p-value = 3.42e-10, total deletions n = 43). For more details, see Fig. 3E. **f**. Table shows the microhomology results of the re-analysis of the public *C. elegans* and human genome sequencing datasets. The heatmap plots are shown in Extended Data Fig. 7 a-e. Binomial test was used, p-values are shown.

Extended Data Fig. 8 HIS-24 regulates the paternally inherited embryonic lethality.

a. RNAi screening of ten DNA-associated protein candidates in hermaphrodite WT animals for the paternally inherited transgenerational embryonic lethality. Knockdown of *his-24* significantly reduced the progeny lethality in the F1 generation. n = 3 biological replicates, medians with 95% CI are shown (p-values are shown. GLM with logit link function and Tukey multiple comparisons were used). **b**. Progeny lethality characterization of *EV (RNAi); fog-2* and *his-24 (RNAi); fog-2* with or without paternal IR irradiation (60 Gy and 90 Gy). n indicating biological replicates are shown. Medians with 95% CI are shown (p-values are shown. GLM with logit link function was used). **c**. Progeny lethality characterization of *EV (RNAi); fog-2*, *his-24 (RNAi); fog-2*, and *hpl-1 (RNAi); fog-2* with or without paternal IR irradiation (90 Gy). n indicating biological replicates are shown. Medians with 95% CI are shown (p-values are shown, GLM with logit link function and Tukey multiple comparisons were used). **d**. Knockdown of *EV, his-24, hpl-1* and *hpl-2* in hermaphrodite WT and *brc-1* mutants with parental IR exposure (90 Gy) at the late L4 stage, and characterization of the progeny lethality in the F1 generation. Medians with 95% CI are shown (n indicating biological replicates are shown, p-values are shown. GLM with logit link function and Tukey multiple comparisons were used). One-way ANOVA (**a** and **c**) Two-way ANOVA (**d**) and two-tailed t-test (**b**) with arcsine transformation were also used for the proportion data to confirm the statistical results and are included in the full statistical results shown in Supplementary table 1.

Supplementary information

Reporting Summary

Supplementary Table 1

Complete statistical analyses.

Supplementary Table 2

Number of the new translocations in the F_1 offspring per sample.

Supplementary Table 3

Complete list of called chromosomal structural variation sites before and after filtering with P_0.

Supplementary Table 4

The results of the proteomic GSEA for chromosomal enrichment.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Wang, S., Meyer, D.H. & Schumacher, B. Inheritance of paternal DNA damage by histone-mediated repair restriction. *Nature* **613**, 365–374 (2023). https://doi.org/10.1038/s41586-022-05544-w

- Received: 27 April 2022
- Accepted: 08 November 2022
- Published: 21 December 2022
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05544-w

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

This article is cited by

- **A mechanism for inheriting radiation-induced DNA damage**
  - Ronald Cutler
  - Jan Vijg

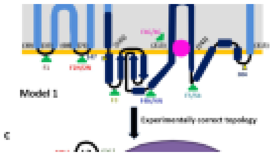
Nature (2022)

A mechanism for inheriting radiation-induced DNA damage

- Ronald Cutler
- Jan Vijg

Nature News & Views 21 Dec 2022

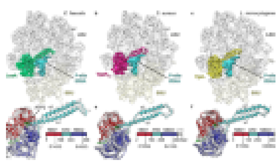
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05544-w



An experiment-informed signal transduction model for the role of the Staphylococcus aureus MecR1 protein in β-lactam resistance

20 December 2019

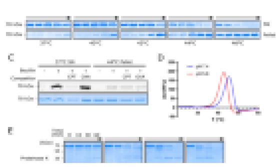
Bruno S. Belluzo, Luciano A. Abriata, … Leticia I. Llarrull



Structural basis of ABCF-mediated resistance to pleuromutilin, lincosamide, and streptogramin A antibiotics in Gram-positive pathogens

11 June 2021

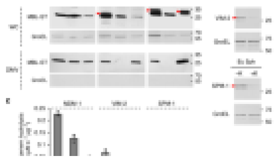
Caillan Crowe-McAuliffe, Victoriia Murina, … Daniel N. Wilson



Thermosensitive PBP2a requires extracellular folding factors PrsA and HtrA1 for Staphylococcus aureus MRSA β-lactam resistance

15 November 2019

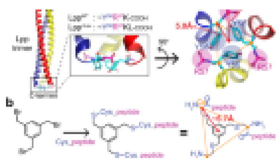
Mélanie Roch, Emmanuelle Lelong, … William L. Kelley



[Protein determinants of dissemination and host specificity of metallo-β-lactamases]

09 August 2019

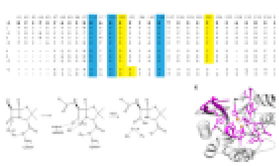
Carolina López, Juan A. Ayala, … Alejandro J. Vila



[Over-activation of a nonessential bacterial protease DegP as an antibiotic strategy]

01 October 2020

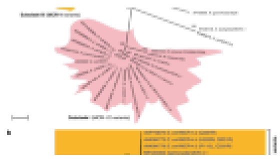
Hyunjin Cho, Yuri Choi, … Seokhee Kim



[Differential active site requirements for NDM-1 β-lactamase hydrolysis of carbapenem versus penicillin and cephalosporin antibiotics]

30 October 2018

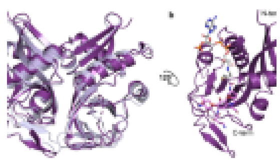
Zhizeng Sun, Liya Hu, … Timothy Palzkill



Action and mechanism of the colistin resistance enzyme MCR-4

25 January 2019

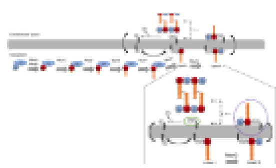
Huimin Zhang, Mengyun Hou, … Youjun Feng



Structural and phylogenetic analyses of resistance to next-generation aminoglycosides conferred by AAC(2′) enzymes

02 June 2021

Angelia V. Bassenden, Linda Dumalo, … Albert M. Berghuis



New insights into the resistance mechanism for the BceAB-type transporter SaNsrFP

10 March 2022

Julia Gottstein, Julia Zaschke-Kriesche, … Sander H. J. Smits

- Article
- Open Access
- Published: 04 January 2023

Structural basis of broad-spectrum β-lactam resistance in *Staphylococcus aureus*

- J. Andrew N. Alexander,
- Liam J. Worrall,
- Jinhong Hu,
- Marija Vuckovic,
- Nidhi Satishkumar,
- Raymond Poon,
- Solmaz Sobhanifar,
- Federico I. Rosell,
- Joshua Jenkins,
- Daniel Chiang,
- Wesley A. Mosimann,
- Henry F. Chambers,
- Mark Paetzel,
- Som S. Chatterjee &
- …
- Natalie C. J. Strynadka

Show authors

Nature **volume 613**, pages 375–382 (2023)

- 11k Accesses
- 97 Altmetric
- Metrics details

Abstract

Broad-spectrum β-lactam antibiotic resistance in *Staphylococcus aureus* is a global healthcare burden[1,2]. In clinical strains, resistance is largely controlled by BlaR1[3], a receptor that senses β-lactams through the acylation of its sensor domain, inducing transmembrane signalling and activation of the cytoplasmic-facing metalloprotease domain[4]. The metalloprotease domain has a role in BlaI derepression, inducing *blaZ* (β-lactamase PC1) and *mecA* (β-lactam-resistant cell-wall transpeptidase PBP2a) expression[3,4,5,6,7]. Here, overcoming hurdles in isolation, we show that BlaR1 cleaves BlaI directly, as necessary for inactivation, with no requirement for additional components as suggested previously[8]. Cryo-electron microscopy structures of BlaR1—the wild type and an autocleavage-deficient F284A mutant, with or without β-lactam—reveal a domain-swapped dimer that we suggest is critical to the stabilization of the signalling loops within. BlaR1 undergoes spontaneous autocleavage in *cis* between Ser283 and Phe284 and we describe the catalytic mechanism and specificity underlying the self and BlaI cleavage. The structures suggest that allosteric signalling emanates from β-lactam-induced exclusion of the prominent extracellular loop bound competitively in the sensor-domain active site, driving subsequent dynamic motions, including a shift in the sensor towards the membrane and accompanying changes in the zinc metalloprotease domain. We propose that this enhances the expulsion of autocleaved products from the active site, shifting the equilibrium to a state that is permissive of efficient BlaI cleavage. Collectively, this study provides a structure of a two-component signalling receptor that mediates action—in this case, antibiotic resistance—through the direct cleavage of a repressor.

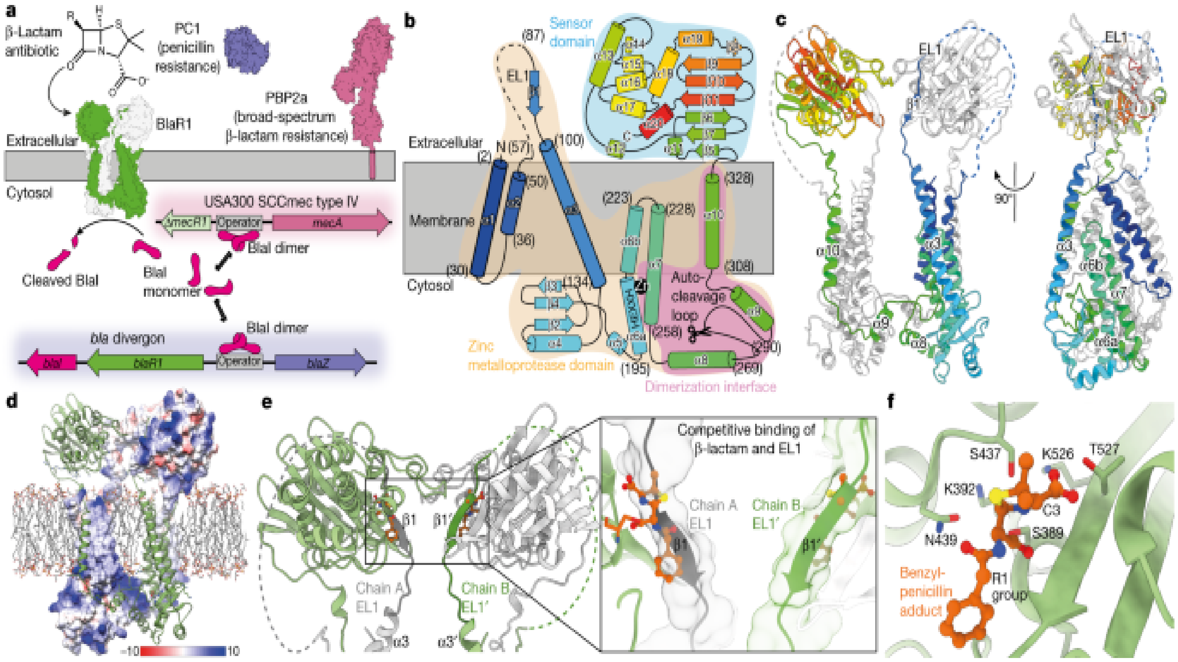
Main

Staphylococcus aureus underlies one of the top three most deadly drug-resistant bacterial infections worldwide[2]. Strains of methicillin-resistant *S. aureus* (MRSA) are particularly problematic owing to their prevalence in nosocomial and community-acquired infections, increased rates of morbidity and death, as well as elevated healthcare costs[1]. MRSA is notorious for its ability to acquire resistance to not only β-lactams but also many other classes of antibiotics[1,9]. The drugs typically used to treat MRSA infections, such as vancomycin, are less effective, more toxic and are subject to

established and growing resistance mechanisms as evidenced in vancomycin-resistant *S. aureus* strains[9,10,11]. Despite some progress in the development of new therapeutics[1], *S. aureus* has recently been estimated to cause more than a quarter of all mortality attributed to antimicrobial resistance in economically developed countries[2], demonstrating that improved treatment options are urgently required. A clearer understanding of resistance in this pathogen could facilitate the development of new cocktail therapies to preserve β-lactam action.

Inducible β-lactam resistance in *S. aureus* is mediated by the homologous *bla* or *mec* divergons. These encode regulatory systems that sense environmental β-lactams through a membrane-embedded sensor–inducer (BlaR1/MecR1; 35% sequence identity; Extended Data Fig. 1) and induce cleavage of a cytoplasmic transcriptional repressor (BlaI/MecI; 61% sequence identity)[7]. This results in the expression of *blaZ* and *mecA*, which encode β-lactamase PC1[5] (conferring resistance to penicillin antibiotics) and the altered penicillin-binding protein (enzymes involved in cell-wall synthesis) PBP2a (conferring insensitivity to almost all other β-lactams)[6,12], respectively (Fig. 1a). The presence of *mec*, embedded in a mobile genetic element called the Staphylococcal cassette chromosome (SCC*mec*), and the production of PBP2a are hallmarks of MRSA[13]. However, there is a complex interplay between divergons. BlaR1–BlaI can control *mecA* expression[3,14,15] (Fig. 1a), notably inducing PBP2a production in response to β-lactam in minutes compared with hours with MecI–MecR1, which can leave the bacterium susceptible to drug treatment[16]. Accordingly, many SCC*mec* types lack functional MecR1 and/or MecI[17]. One of the predominant MRSA strains in the United States is clonal complex 8 USA300 in which *mecA* is located on a type IV SCC*mec* cassette with a truncated *mecR1* and a deleted *mecI*[17,18] with expression under the control of plasmid-encoded BlaR1 and BlaI. More generally, type IV SCC*mec* or its closely related subtypes are frequently encountered owing to their presence in epidemic community MRSA strains worldwide. Thus, BlaR1 and BlaI are critically important in determining broad-spectrum antibiotic resistance in many MRSA strains.

Fig. 1: The cryo-EM structure of *S. aureus* BlaR1.



a, The *bla* pathway controlling β-lactam antibiotic resistance in *S. aureus*. Previously published models for PC1 (Protein Data Bank (PDB): 3BLM) and PBP2a (PDB: 1MWS) were used to generate the figure. **b**, Topology of BlaR1. The zinc metalloprotease and sensor domains are highlighted, as well as the dimerization interface. Residue numbers are indicated in parentheses. **c**, Ribbon representation of the BlaR1 dimer with the monomers shown in either sequence (N to C) rainbow or white. **d**, Dimeric BlaR1 embedded in a lipid bilayer (Methods). Each protomer of the domain-swapped dimer is shown either as a coulombic electrostatic surface (kcal (mol e^{-})$^{-1}$ at 298 K) or a green ribbon. Positively charged residues (Arg, Lys, His) proximal to membrane are shown as sticks. SXXK (sensor) and HEXXH (protease) motif residues are shown in ball and stick representation. **e**, Ribbon model of the BlaR1 dimer coloured white for chain A and light green for chain B. BlaR1 EL1 is shown in dark grey and dark green for chains A and B, respectively. A superposed benzylpenicillin molecule is shown in orange from an aligned X-ray structure of the BlaR1 sensor domain (PDB: 1XA7). Inset: EL1 as both a ribbon and a transparent solvent-excluded surface, demonstrating that the modelled position of EL1 occupies the β-lactam-binding cleft and would be competitive with the superposed benzylpenicillin antibiotic. **f**, X-ray structure of the *S. aureus* BlaR1 sensor-domain active site with covalently bound benzylpenicillin (PDB: 1XA7). Residues of key catalytic motifs are shown as sticks.

BlaR1 and MecR1 are membrane proteins with an N-terminal zinc metalloprotease domain and a C-terminal extracellular β-lactam-binding sensor domain. Isolated structures of the β-lactam-binding sensor domain in complex with antibiotics have been determined for *S. aureus* BlaR1 and MecR1, showing near-identical architectures (Cα root mean squared deviation of around 2.9 Å) that are related to the class D serine β-lactamases[15]. No structures of the transmembrane and proteolytic regions have been determined and the signal transduction events linking β-lactam binding and BlaI cleavage on opposing sides of the membrane remain unclear[19,20,21]. To address these unknowns, we present structures of full-length *S. aureus* BlaR1. On the basis of conformational differences with and without the β-lactam ampicillin, we propose a mechanism of *bla* regulation that defines BlaR1 as a two-component signalling receptor. These insights provide a molecular foundation for the development of therapeutics for both MRSA and homologous pathways in other drug-resistant pathogens, including *Mycobacterium tuberculosis*[22] and *Clostridioides difficile*[23].

Cryo-EM analysis of BlaR1

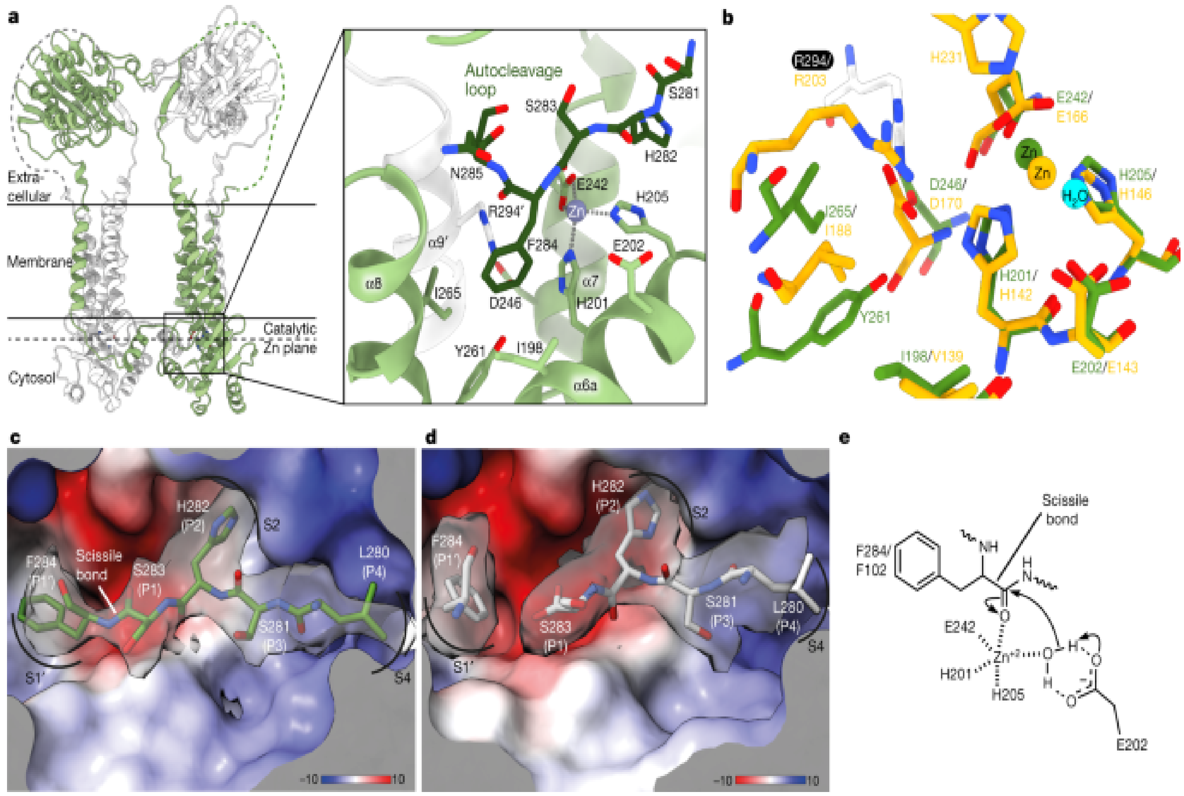
To overcome the historical challenge of obtaining adequate quantities of active BlaR1[21,24], we used the nisin-controlled gene expression system in *Lactococcus lactis*[25]. Membranes containing expressed wild-type BlaR1 (BlaR1(WT)) were able to cleave BlaI, suggesting that the zinc metalloprotease domain is active (Extended Data Fig. 2a), as shown previously in *Escherichia coli* membranes containing expressed *S. aureus* BlaR1[26]. Purification of BlaR1 was monitored by the covalently bound fluorescent β-lactam BOCILLIN FL[27]. As in *S. aureus* and *E. coli*[28], BlaR1(WT) expressed in *L. lactis* undergoes spontaneous autocleavage (Extended Data Fig. 2c) between residues Ser283 and Phe284 (confirmed by Edman N-terminal sequencing), which was mitigated by mutation of Phe284 in the P1′ position to alanine (BlaR1(F284A)) (Extended Data Fig. 2e). Detergent-solubilized BlaR1(WT) and BlaR1(F284A) behave like oligomers (Extended Data Fig. 2f) and rate-zonal ultracentrifugation was used to enrich the sample for the intact oligomeric species (Extended Data Fig. 2b–e).

The structures of BlaR1(WT) and BlaR1(F284A) (the latter alone and in complex with ampicillin (Extended Data Fig. 2e)) were determined using cryo-electron microscopy (cryo-EM) to nominal resolutions of 3.8–4.6 Å (Fig. 1, Extended Data Figs. 3–5 and Extended Data Table 1). BlaR1 was consistently observed in a dimeric form (Extended Data Fig. 3b,c) with two-fold symmetry averaging applied in the overall reconstructions for each sample. More structural heterogeneity was detected in the BlaR1(WT) dataset compared with the uncleavable mutant BlaR1(F284A), suggesting enhanced dynamic motions. An asymmetric homodimer class was identified, resulting in a reconstruction of the previously uncharacterized BlaR1(WT) transmembrane and zinc metalloprotease region to 3.8 Å with a local resolution extending to about 3.3 Å (Extended Data Fig. 3e).

BlaR1 forms a domain-swapped dimer

Notably, BlaR1 is observed as an extensive and intimate dimer mediated by an unambiguous domain swap (Fig. 1c,d, Extended Data Fig. 6e and Supplementary Video 1). The N-terminal zinc metalloprotease domain is well-ordered and includes four full transmembrane α-helices, a re-entrant loop (often implicated in meditating protein channel specificity, transport or signalling functions)[29] and a cytosolic zinc metalloprotease active site defined by the characteristic H_{201}EXXH and E_{242}XXXD gluzincin signature motifs[30] (Figs. 1b and 2 and Extended Data Fig. 1). The active site projects towards the cytosolic face, allowing a point of access for BlaI binding, while the soluble C-terminal β-lactam-sensing domain protrudes from the opposing extramembranal face of the bilayer (Figs. 1 and 2). The N terminus is also oriented to the extracellular side of the membrane leading to an N_{out}, C_{out} topology (Fig. 1b,c), which is rarely observed in bacterial membrane proteins[31,32].

Fig. 2: The zinc metalloprotease active site of BlaR1 and binding of the autocleavage loop substrate and product.



a, The BlaR1 dimer is shown as a ribbon and is coloured by monomer. The position of the bound catalytic zinc (dashed line) relative to the membrane leaflet (solid line) is indicated. The box out shows the labelled zinc metalloprotease active-site residues, highlighting the intermolecular composition (residues are shown in stick with CPK colouring, with carbons in light green, except for those of the domain-swapped contributing residue Arg294, which are shown in white). Autocleavage loop residues are shown in darker green. **b**, Alignment of BlaR1 zinc metalloprotease active-site residues (depicted as described in **a**) with thermolysin (PDB: 2TLX), shown in CPK with gold carbons. The proposed thermolysin catalytic water is represented as a cyan sphere. **c**,**d**, Electrostatic surface representation (kcal $(\text{mol e}^{-})^{-1}$ at 298 K) of the BlaR1(WT) active-site cleft with intact (**c**) and cleaved (**d**) autocleavage loop residues, shown in CPK colouring with green and white carbons, respectively. The cryo-EM map is shown as a transparent grey surface. See also Extended Data Fig. 4b,c. **e**, A 2D representation of the proposed steps leading to the formation of a oxyanion tetrahedral transition state in the zinc metalloprotease active site.

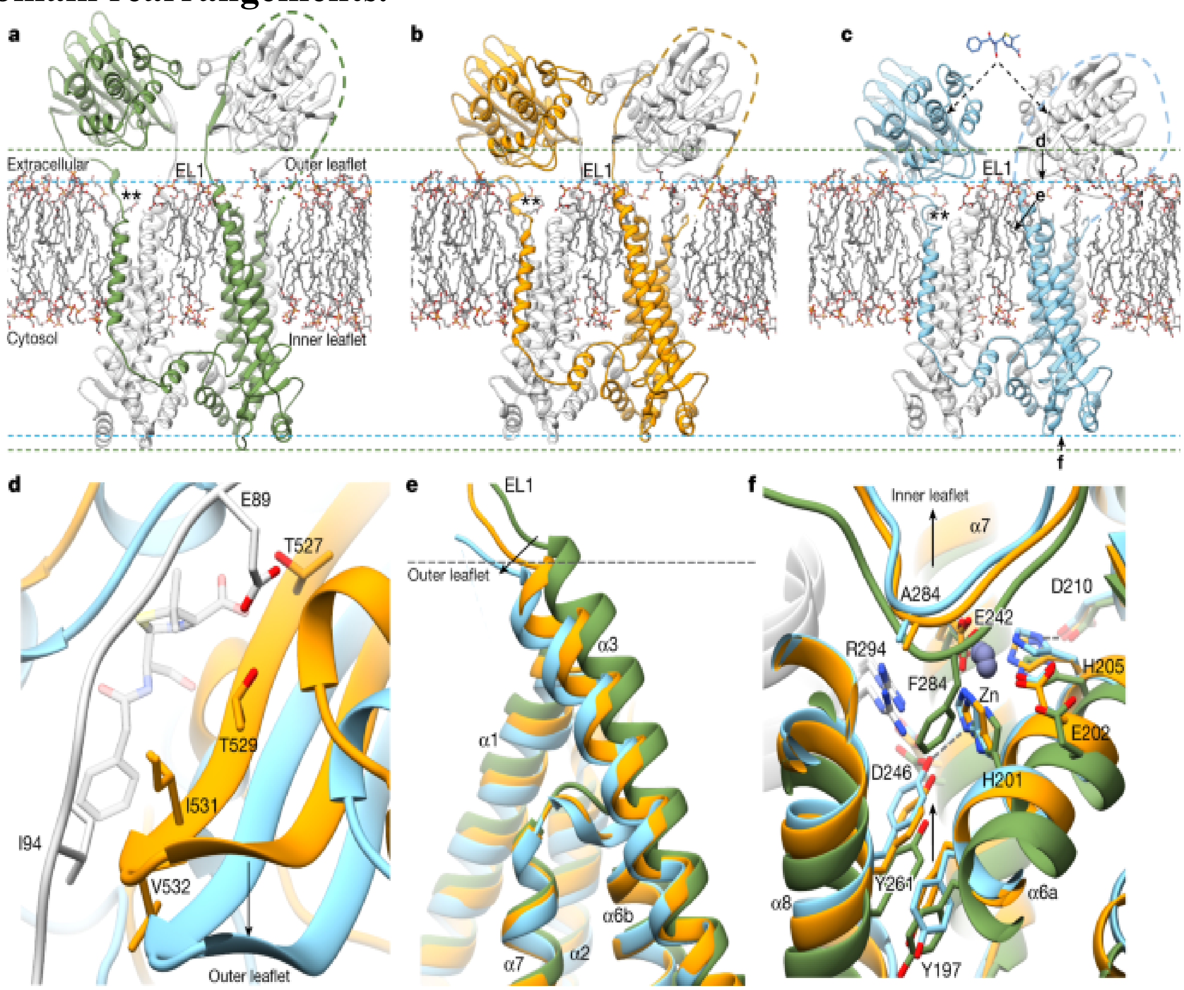
The domain swap involves the cytosolic-facing helix α9 of the zinc metalloprotease domain and helix α10, the terminal transmembrane helix, resulting in an extensive interface with a buried surface area of about 3,165 Å^2. Within the zinc metalloprotease domain, additional interface interactions involve helices α8 (same chain) and α9′ (domain swapped), which are oriented almost parallel to the membrane, and helix α7, the C-terminal re-entrant helix that houses the H_{201}EXXH motif. Dimerization defines a central cavity, which in the native membrane is presumably filled with lipids (Fig. 1d). Densities consistent with phosphatidylglycerol headgroups are observed lining the electropositive surface formed at the base of the cavity. Additional diffuse density for fatty acid chains enabled modelling of phosphatidylglycerol(16:0/16:0), a common lipid found in *L. lactis* and *S. aureus*[33,34,35,36] (Extended Data Fig. 6a–c).

The loop connecting helices α8 and swapped α9, housing the scissile bond between residues Ser283 and Phe284 (residues 271 to 289; here termed the autocleavage loop), enters the zinc metalloprotease active site of its own chain, directly demonstrating that autocleavage occurs in *cis* (Fig. 2a). The swapped position of α9 effectively lengthens the distance between the autocleavage loop tether points and creates a stabilizing set of hydrophobic core contacts at the dimeric interface involving a number of conserved side chains. In the BlaR1(F284A) mutant structures, the autocleavage loop is intact; however, in the BlaR1(WT) structure, we detected a mixture of uncleaved and cleaved forms, the latter with the central P1′ Phe284 slightly shifted in position (Fig. 2c,d, Extended Data Fig. 4b,c and Supplementary Video 2). This observation is consistent with the presence of both cleaved and uncleaved BlaR1(WT) observed using SDS–PAGE (Extended Data Fig. 2c). We propose that dimerization results in encapsulation of the autocleavage loops within the lumen of the dimer, potentially providing enhanced stability of these dynamic signalling regions to unbridled conformational freedom and/or random hydrolysis.

The extracellular BlaR1 sensor domain was less-well resolved, suggesting a more dynamic relative orientation. The reconstructions of BlaR1(WT) and BlaR1(F284A) without ampicillin provided sufficient secondary structure features (local resolution 6–6.5 Å) to enable confident placement of the sensor-domain X-ray crystal structure[37] (Extended Data Figs. 6d,f and 7a–

c), which is oriented with the β-lactam-binding grooves facing each other (Fig. 1c–e). In the BlaR1(WT) dataset, the sensor domains were observed to be dynamic and are able to adopt different orientations with respect to each other and the transmembrane domains (Fig. 3a, Extended Data Fig. 7 and Supplementary Video 3). A comparison with BlaR1(F284A) shows that the sensor domains are positioned at different distances to the transmembrane domain in order of BlaR1(WT), BlaR1(F284A) and BlaR1(F284A)–ampicillin (Fig. 3a–c and Extended Data Fig. 7a–c).

Fig. 3: Conformational changes after β-lactam binding to the extracellular sensor domain, resulting in EL1 displacement, sensor reorientation, and transmembrane and cytosolic-facing metalloprotease domain rearrangements.



a–**c**, Structures of dimeric BlaR1(WT) (**a**; green/white ribbon), BlaR1(F284A) (**b**; orange/white ribbon) and BlaR1(F284A)–ampicillin (**c**;

cyan/white ribbon). The dashed lines highlight the extreme positions of the sensor (extracellular) or zinc metalloprotease (cytosol) domains with respect to the membrane and are coloured according to the BlaR1(WT) or BlaR1(F284A)–ampicillin structures. The labelled arrows in **c** highlight the β-lactam-induced movement towards the outer and inner leaflets as expanded on in **d**–**f**. Unresolved EL1 residues 56–86 are indicated by dashed lines. The double asterisk indicates the location of the determined signal peptidase cleavage site Gly331. **d**, Benzylpenicillin from PDB 1XA7 (transparent grey) superposed onto **b** (BlaR1(F284A), orange), showing the structural overlap with the modelled EL1 loop (white). β-Lactam binding as described in **c** (BlaR1(F284A)–ampicillin, cyan) induces a shift in the sensor domain towards the outer leaflet of the membrane, we propose as a result of competitive displacement of EL1 from the sensor-domain active site. **e**, β-Lactam-induced motions of transmembrane helix α3, directly downstream of EL1, coloured as described in **a**–**c**. The distance between the sensor domain and the membrane correlates with increased displacement of the N-terminal region of α3. **f**, Superposed active-site structures from BlaR1 structures in **a**–**c**. The F284A mutation results in an unoccupied S1′ pocket and associated movement of Tyr261 on helix α8 towards His201 of the HEXXH motif in **b** (orange) and **c** (cyan; dashed line). Supporting data are provided in Extended Data Fig. 7. Superpositions in **d**–**f** are based on structural alignment of the re-entrant helices containing the catalytic HEXXH…E motif.

The extracellular loop (EL1; residues 56–96), which has long been hypothesized to be involved in BlaR1 signalling[19,20], was observed to protrude from transmembrane helices α2 and α3 at the outer leaflet into the region adjacent to the sensor domain of the neighbouring monomer as a result of the domain swap (Fig. 1c–e). Like the sensor domains, the EL1 loops are less-well resolved, consistent with a more dynamic nature. EL1 was modelled using AlphaFold2[38] as a starting template and optimized using Rosetta density-guided refinement[39] (Fig. 1e, Extended Data Fig. 6g–i and Supplementary Video 1). Conserved residues Ile87–Ile94 were specifically observed occupying the β-lactam-binding groove in *trans*. Notably, acidic Glu89 (consensus Asp) and hydrophobic Ile94 (consensus Val) occupy a similar space to the C3 β-lactam carboxylate and hydrophobic R1 groups in structures of isolated acylated sensor domains (Fig. 1e,f and Extended Data Fig. 6i). The β-lactam carboxylate forms a near universal favourable

electrostatic interaction with the adjacent electropositive KTG motif (Lys526 in BlaR1) and the R1 group forms extensive van der Waals interactions with the edge strand in the active site (BlaR1 residues Gly530–Val532). In support of these competitive binding interactions, a peptide with the sequence of EL1 has previously been found to interact with the sensor domain in solution[19,20,40]. It is clear that acylation of the sensor-domain active site serine by β-lactam antibiotics will displace the EL1 loop from the active site, a structural hallmark of initiation of downstream signalling events now illuminated here.

The protease active site and cleavage mechanism

Helices α6, α7 and β-strand β3 together take on a thermolysin-like fold (the prototypical gluzincin), protruding from a transmembrane scaffold, with analogous positioning of several key active site residues (Fig. 2b). Readily identified by localization of the gluzincin catalytic metalloprotease HEXXH…E sequence motif[41], BlaR1 helix α6a here positions His201 and His205, along with Glu242 from neighbouring helix α7, to coordinate the clearly observed catalytic zinc in the expected tetrahedral manner (Fig. 2a and Extended Data Figs. 4b,c and 7e) and present at a 1:1 molar ratio with BlaR1 (Supplemental Table 1). Helices α8 (specifically Ile265 in the S1′ subsite and Tyr261 approaching His201) and α9′ (Arg294′, which forms a stabilizing interchain salt bridge with Asp246) both contribute conserved residues to the active site, suggesting that their precise position is important for function and supporting the observed domain-swapped dimer[42] (Fig. 2a,b).

The initial zinc metalloprotease substrate, the autocleavage loop captured here in uncleaved and cleaved states (Fig. 2c,d), provides insights into the molecular details of substrate specificity. The P4–P1′ residues of the autocleavage loop (Leu280, Ser281, His282, Ser283, Phe284) adopt the typical β-strand conformation of protease substrates[43], which runs anti-parallel to β3 of the gluzincin domain β-sheet analogous to substrate binding in thermolysin related proteases (Fig. 2). In addition to main-chain β-strand hydrogen bonds between Thr172 and His282, the substrate is anchored by hydrophobic interactions of Leu280 into the S4 pocket at one end and P1′ Phe284 inserting into the S1′ subsite at the other, with the intervening P2

His282 within the S2 pocket the other important specificity determinant. The well-defined S1′ subsite, typically the major determinant for substrate specificity in endolytic metalloproteases[44], has a large hydrophobic cavity that is suited for the binding of the aromatic phenylalanine side chain of P1′ Phe284 (Fig. 2c,d). Sequence conservation at this position shows that there is a preference for a bulky hydrophobic amino acid (Extended Data Fig. 8b). By contrast, phenylalanine at the P1′ of the BlaI repressor cleavage sequence is almost invariant (Extended Data Fig. 8c), hinting at a tighter affinity that may be important in displacing autocleaved products from the BlaR1 metalloprotease active site.

In the WT structure, one chain shows clear separation of Ser283(P1) and Phe284(P1′) supporting a post-autocleavage state (Fig. 2c,d and Extended Data Fig. 4b,c). By contrast, the density in the alternate chain is more uninterrupted, suggesting a pre-cleavage state, consistent with a population of both uncleaved and cleaved protein in the sample used for cryo-EM. The autocleavage loop is in the approximate position for cleavage, with the P1 Ser283 carbonyl orientated favourably towards the catalytic zinc, which canonically acts to polarize the scissile bond, promoting its electrophilicity and subsequent nucleophilic attack by an activated water to promote cleavage. Although we observed the same binding orientation in the autocleavage-deficient BlaR1(F284A) structure, the P1′ alanine mutant leads to the loss of the close van der Waals-driven socketing of the aromatic phenylalanine side chain into the S1′ pocket, preventing productive coordination of the carbonyl oxygen to the zinc. Although the hydrolytic water was not observed at this resolution, by comparison with thermolysin, we expect that Glu202 in BlaR1 coordinates the catalytic water and provides general base activation to the nucleophilic form for attack on the P1 carbonyl and subsequent hydrolysis (Fig. 2b,e).

Notably, both the intact and cleaved autocleavage loop forms occupy the catalytic cleft such that their presence is not sterically compatible with the binding and cleavage of BlaI. Thus, our structures show that movement of the autocleavage loop out of the active site, either in an intact or cleaved state, is needed for cleavage of BlaI to proceed. We suggest again a role for the observed dimerization in encompassing and stabilizing these loops in both forms at the dimeric interface, minimizing their susceptibility to

random proteolysis that would presumably be enhanced in the more exposed, even less stable monomeric forms.

BlaR1 autocleavage is spontaneous

The role of autocleavage in the regulation of BlaR1 catalytic activity and in turn the BlaR1 divergon has been an important focus[4,26,28]. Here, BlaR1(WT) undergoes spontaneous autocleavage (Extended Data Fig. 2c), consistent with previous reports[4,26,28]. Autocleavage is time and temperature dependent with substantial albeit incomplete cleavage at 4 °C, which was used for cryo-EM sample preparation, and near-complete fragmentation overnight at room temperature or 37 °C (Extended Data Fig. 2g).

To further investigate the role of autocleavage, we compared the growth of BlaR1(WT) and BlaR1(F284A) in the common clinical variant USA300 MRSA strain SF8300, in which *mecA* expression is under control of a *blaI-blaR1* plasmid, in the presence of the β-lactam antibiotic nafcillin (Extended Data Fig. 8a). In the absence of antibiotic, all of the strains displayed a similar growth pattern. Nafcillin treatment fully inhibited the growth of the strain containing *blaR1(null)-blaI*, presumably due to its inability to derepress BlaI. The strains containing an empty vector or a non-functional *blaR1(null)-blaI(null)* construct did not produce growth inhibition, suggesting constitutive *mecA* expression in absence of a functional BlaI repressor. The strain containing BlaR1(F284A) (*blaR1(F284A)-blaI*) displayed intermediate growth attenuation in presence of nafcillin compared with either wild-type BlaR1 (*blaR1-blaI*) or the functionally defective BlaR1 (null) (*blaR1(null)-blaI*) (Extended Data Fig. 8a), showing that the F284A mutant is partially functional.

We further examined the ability of purified BlaR1—including both the variants used in our structural studies and that used in the growth assay above (BlaR1(USA300))—to directly cleave purified BlaI. BlaR1(WT) and BlaR1(USA300) cleave BlaI in a manner that is completely abrogated by EDTA but not serine/cysteine protease inhibitors, supporting specific metalloprotease activity (Extended Data Fig. 8d–f). Edman N-terminal sequencing confirmed cleavage between BlaI residues Asn101 (P1) and Phe102 (P1′), as previously identified in BlaR1-containing *S. aureus*

membranes[4]. Autocleavage-deficient BlaR1(F284A) also shows BlaI cleavage, albeit at a reduced level compared with BlaR1(WT), suggesting that some level of BlaI binding and cleavage is still possible in the presence of an intact autocleavage loop (barring a minor population of autocleaved BlaR1(F284A) that is not detectable in our fluorescence-based gel assay (Extended Data Fig. 2e).

Collectively, these results suggest that the autocleavage event is spontaneous and is not part of the signal transduction mechanism itself, and it does not appear to be an absolute requirement for BlaI cleavage and activation of the *bla* divergon to occur, although it does permit a more efficient response to environmental β-lactams that may be required for resistance in vivo. These data also show that BlaR1 can recruit and cleave BlaI directly with no need for intermediary components, as was proposed previously[8].

β-Lactam-induced allosteric changes

β-Lactam binding to the extracellular BlaR1 sensor domain regulates BlaI cleavage by the zinc metalloprotease domain in the cytoplasm. The three structures presented here reveal conformational changes linking the two BlaR1 active sites about 65 Å apart across the intervening membrane. The overall transmembrane and zinc metalloprotease domains are similar across structures suggesting that no major conformational rearrangements are induced on acylation, although subtle motions of select helices were observed that we believe are functionally important. By contrast, we observed that the sensor domain is more dynamic (Fig. 3a). Focused 3D classification shows that the BlaR1(WT) sensor domain samples multiple orientations, although none refine to a high resolution, which indicates a continuous dynamic trajectory. 3D variability analysis[45] shows a general up/down translation of 4–5 Å relative to the membrane (Supplementary Video 2). The BlaR1(F284A) structure (refined with C_2 symmetry) exhibited less variability and the sensor domains adopt a position slightly closer to the membrane compared with BlaR1(WT) (Fig. 3a–c and Extended Data Fig. 7a–c). However, in the presence of ampicillin, we see a pronounced rigid body shift of about 6 Å to a position immediately abutting the membrane. The magnitude of this shift is consistent with the expected displacement of EL1 induced by the competitive binding of ampicillin (Fig. 3d). Thus, β-

lactam binding and the consequent extrusion of EL1 from the active site appears to trigger movement of the sensor domain towards the outer leaflet of the membrane.

Correlated motions are also observed in the transmembrane and zinc metalloprotease domains. Notably, a general compression parallel to the membrane normal was observed in the BlaR1(F284A)–ampicillin structure. Careful superposition of the reconstructions onto the re-entrant helices containing the catalytic HEXXH…E motif clearly demonstrates that structural elements on either side of the membrane are compressed towards each other in the presence of ampicillin (Fig. 3a–c and Extended Data Fig. 7). As above, a key determinant of the observed motions appears to be the steric displacement of EL1 by the shifted sensor domain. EL1 is situated immediately upstream of helix α3, which is observed kinked into the outer leaflet of the membrane as a result of the shift of EL1 (Fig. 3e and Extended Data Fig. 7d). This is coupled to a general movement of the distal region of the zinc metalloprotease domain towards the inner leaflet of the membrane (Fig. 3a–c and Supplementary Video 3).

Discussion

The structure of *S. aureus* BlaR1 has been unclear since the initial identification of the gene over 30 years ago[3,46]. This paper advances knowledge of BlaR1-mediated antibiotic resistance, providing insights into the mechanism of BlaR1 signal transduction that could facilitate the development of therapeutics to treat MRSA infections.

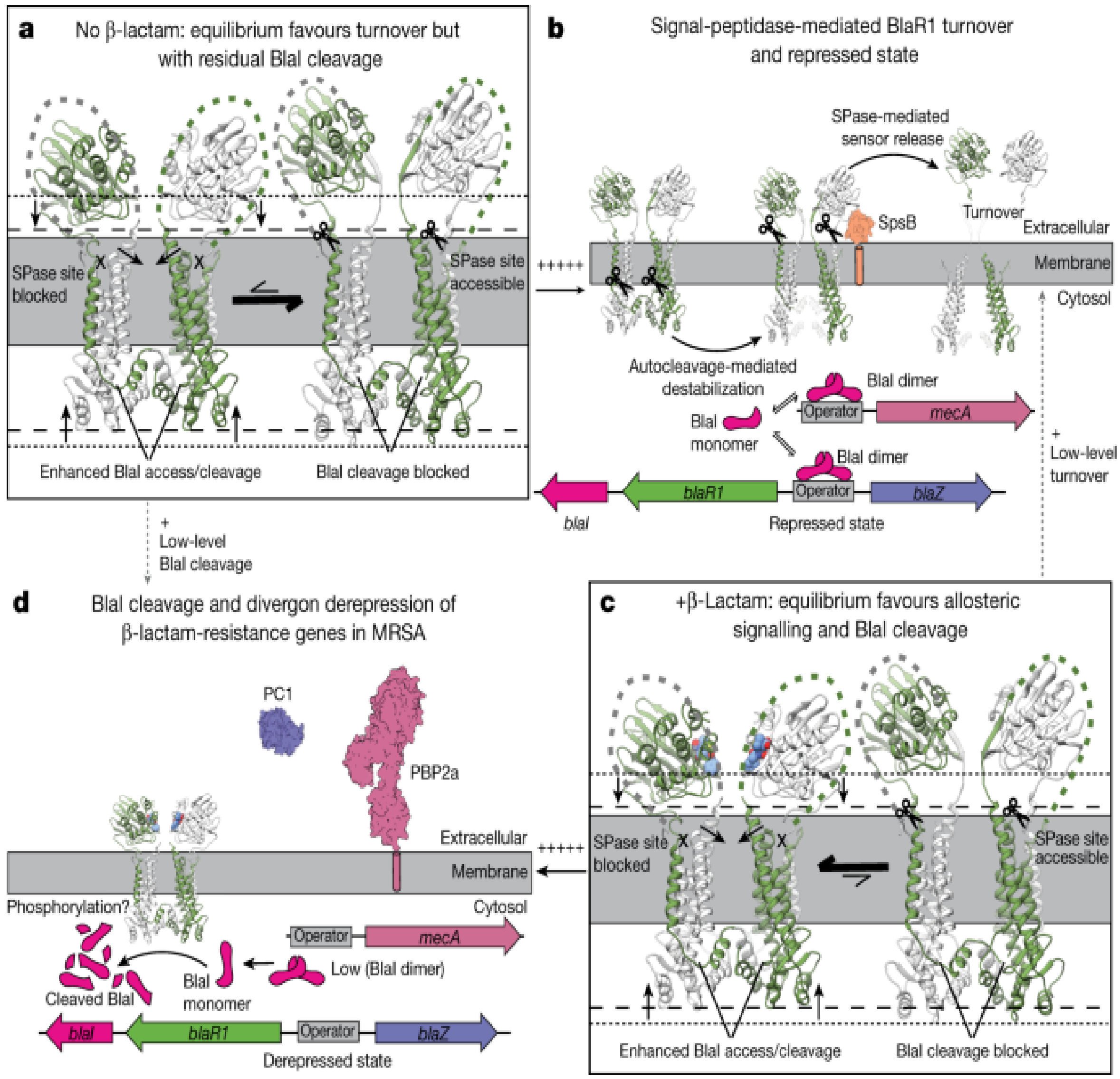
BlaR1 forms a domain-swapped dimer, with the zinc metalloprotease active site accessible from the cytoplasm. The gluzincin fold superposes well with the prototypical soluble representative thermolysin and the most closely related integral membrane zinc metalloproteases with known structures: eukaryotic nuclear membrane ZMPSTE24[47] and Ste24p[48]. In addition to the conserved H_{201}EXXH and E_{242}XXXD motifs, the analysis further defines residues that are critical to catalytic function. Asp210 and Tyr261 are near strictly conserved and are positioned to stabilize core catalytic residues His205 or His201, respectively (Fig. 3f and Extended Data Fig. 9). Mutation of either to alanine abolished β-lactamase induction in *B. licheniformis*[49].

Asp246 forms a salt bridge with Arg294′, an interaction that is conserved in other gluzincins; however, in BlaR1, this is located on domain-swapped helix α9′ and forms an intermolecular interaction. In thermolysin, the arginine side chain binds to the substrate P1′ carbonyl[42]. Consistently, Arg294′ projects towards Phe284 of the autocleavage loop in BlaR1. Mutation of Asp246 abolished β-lactamase induction in *B. licheniformis*, whereas double mutation of Arg293 and Arg294 severely diminished activity[49]. We did not observe an equivalent of thermolysin His231, which is proposed to be involved in transition-state binding and stabilization[50] (Fig. 2b). His231 is conserved in other gluzincins, including ZMPSTE24, in which it is located in a loop connecting helices equivalent to BlaR1 α8 and α9, which in BlaR1 is the autocleavage loop.

An outstanding question of BlaR1 function is how sensing environmental β-lactams outside the cell regulates BlaI cleavage inside. The structures here describe concerted motions that link the sensor and metalloprotease domains located on opposite sides of the membrane (Fig. 3 and Supplementary Video 3). We hypothesize that the extremes are representative of changes that are involved in the activation and inhibition of the metalloprotease. Although a range of states exists in all of the samples, the equilibrium is shifted—BlaR1(WT) is dominated by a more extended conformation and the BlaR1(F284A)–ampicillin structure is more compact, with the sensor directly in contact with the outer leaflet of the membrane. In the sensor domain, the substrate-mimicking interactions of EL1 within the β-lactam-binding site and the resulting induced domain movement after displacement explain its previously predicted role in signal transduction[19,20]. On the inner leaflet, concomitant motions are observed, with the metalloprotease domain drawn up to the membrane in the presence of ampicillin. Central to this is helix α3, which traverses the membrane directly linking EL1 and the metalloprotease domain. This helix is bent into the membrane at its N terminus in the BlaR1(F284A)–ampicillin structure, seemingly as a result of the force exerted on EL1 due to the acylation-induced sensor-domain movement (Fig. 3e and Extended Data Fig. 7d). The current resolutions, in particular for the BlaR1(F284A)–ampicillin data, limit analysis of the precise structural changes. However, we note that the overall compression results in a more compact metalloprotease active site with Tyr261 (discussed above) positioned at different distances to His201, moving closer in the

ampicillin-bound state of BlaR1(F284A) or in the post-cleavage state of BlaR1(WT), in which the now-free P1′ Phe284 of the cleaved product is not as deeply socketed in the S1′ subsite (Fig. 3f and Extended Data Fig. 7e). In all of the structures, the protease domain is in an autoinhibited state with the autocleavage loop (both cleaved and uncleaved) occupying the active site, suggesting an equilibrium between active site occupied/autoinhibited and active site accessible states (Fig. 4). We hypothesize that the regulation of BlaI cleavage is dependent on the equilibrium between these conformations, with both spontaneous autocleavage and β-lactam acylation serving to modulate this structural equilibrium, pushing it to one favouring displacement of the cleaved autoinhibitory loop fragments from the active site to allow efficient BlaI repressor access and cleavage after the detection of environmental β-lactam (Fig. 4).

Fig. 4: Conformational equilibrium of BlaR1 governing β-lactam resistance in MRSA.



a, In the absence of β-lactams, BlaR1 primarily exists in the basal state with the sensor and metalloprotease domains in the most extended conformation away from the membrane outer and inner leaflets, respectively. The EL1 loop is bound within the sensor active site, the SPase cleavage site is exposed and the zinc metalloprotease domain is blocked by the autocleavage loop from efficiently cleaving BlaI. The dotted and dashed lines indicate the edge of the sensor and zinc metalloprotease domains in their extended or contracted (with β-lactam) states, respectively. **b**, In this basal BlaR1 state, elevated cytosolic BlaI concentrations favour repressor dimerization and binding to the operator sequence, minimizing the expression of genes encoding resistance proteins. In parallel, SPase cleavage of the sensor domain coupled with autocleavage-mediated destabilization of the

transmembrane region promotes turnover of BlaR1 in the membrane in the absence of a β-lactam threat. **c**, Extracellular β-lactams are detected by the BlaR1 sensor through acylation of the SXXK motif, displacing EL1 and promoting movement of the sensor and the metalloprotease domains towards the membrane (Fig. 3), leading to allosteric activation. These events are thought to stabilize the active state of BlaR1 in the presence of antibiotics by blocking of SPase cleavage and promoting a more catalytically competent metalloprotease active site. Autocleavage loop proteolysis also favours BlaI cleavage, probably due to reduced steric hinderance of BlaI binding. However, autocleavage does not appear to be triggered by β-lactam, and we propose that autocleavage is instead needed for BlaR1 maturation and turnover. **d**, The activated metalloprotease domain cleaves BlaI, decreasing the concentration of BlaI dimer and reducing BlaI binding to the operator. Expression of the divergon genes *mecA* and *blaZ* leads to β-lactam protection by PBP2a (β-lactam insensitive; PDB 4WVJ shown) and PC1 (β-lactamase activity; PDB 3BLM shown), respectively.

S. aureus kinase (Stk1)-mediated phosphorylation of BlaR1 is involved in the activation of the *bla* pathway[51]. Tyrosine phosphorylation was specifically implicated[51], and the only conserved cytoplasmic tyrosine residues are Tyr197, projecting from the C-terminal end of α6a, and the above mentioned Tyr261 on α8 that we observed in a spectrum of active-site-engaged positions (Fig. 3f and Extended Data Fig. 9b). Compellingly, these two side chains stack in an antiparallel manner, with Tyr197 projecting to the exterior and accessible for phosphorylation and Tyr261 towards His201. As such, it is possible that the observed conformational changes may facilitate phosphorylation, inducing added structural changes to destabilize the autocleavage loop products from the active site. Identification of phosphorylated residues and their structural effects could illuminate how phosphorylation contributes to the regulation of resistance and would lend further credence to existing efforts to develop inhibitors targeting Stk1[51].

Another interesting aspect of BlaR1 is the need for turnover of the basal-level-expressed receptor[28]. Signal-peptidase-mediated cleavage of *Staphylococcus epidermidis* BlaR1[52] and shedding of the sensor domain in *S. aureus* cultures at equivalent residues (between Gly331 and Gln332) have been documented[28]. We observed the presence of an analogous ~31 kDa

band, which is acylated by BOCILLIN FL (Extended Data Fig. 2c). The *S. aureus* signal peptidase SpsB cleaves BlaR1(WT), with mutation of BlaR1 Gly331 to lysine abrogating this effect (Extended Data Fig. 8g). This cleavage position sits in the dynamic linker between the C-terminal transmembrane helix α10 and the sensor. In the more extended (minus β-lactam) positioning of the sensor domain away from the membrane (Fig. 3a), this loop would be more accessible for processing by membrane-anchored SpsB. Together with autocleavage, these processing events could promote destabilization of the dimeric interface of BlaR1 on both leaflets and promote inactivation/turnover to prevent potentially deleterious accumulation in the membrane and leaky BlaI cleavage/expression of the resistance genes in the absence of β-lactam. After sensing β-lactam, with reorientation of the sensor domain against the membrane (Fig. 3c), this loop is more protected from processing, maintaining BlaR1 integrity to combat the presence of the bactericidal antibiotic (Fig. 4). These observations provide further support for the ongoing interest in the development of signal peptidase inhibitors, of which the potential for synergistic effects on targets beyond the signal peptide processing roles, such as BlaR1 turnover here, is increasingly compelling.

In summary, the data presented here provide insights into the structure and function of the BlaR1 master controller of β-lactam resistance in *S. aureus*. In addition to clinically important *Staphylococcal* pathogens, analogous signalling systems have been identified in *C. difficile*[23] and *M. tuberculosis*[22]. *C. difficile*—a notoriously persistent spore-forming pathogen—has equivalent components to the *bla* pathway (BlaR1, around 30% sequence identity with *S. aureus*; Extended Data Fig. 1) controlling the expression of a β-lactamase[23]. The *M. tuberculosis* BlaR1 homologue has a similar zinc metalloprotease domain but lacks an equivalent sensor domain making it unclear how β-lactams are detected. One possible functional substitution could involve phosphorylation by PknB, a kinase with a β-lactam-binding PASTA domain that is essential to *M. tuberculosis*[22]. We expect that these BlaR1 homologues have a similar domain-swapped architecture and proteolytic derepression of broad-spectrum resistance mechanisms, making the findings here broadly relevant in terms of mechanistic understanding and structure-guided inhibition for the

preservation of an important drug class for fighting infections—the β-lactam antibiotics.

Methods

Plasmid construction and site-directed mutagenesis

S. aureus BlaR1 (UniProt: Q00419); residues 1–585) or BlaR1 from MRSA strain USA300 (RefSeq: WP_001096381.1; residues 1–585) were cloned into the *Lactococcus lactis* expression vector pNZ8048[53] with a C-terminal hexahistidine tag, using restriction-free cloning[54]. Cloning and site-directed mutagenesis of pNZ8048 was performed in *E. coli* MC1061 grown in LB medium supplemented with 10 μg ml^{-1} chloramphenicol. After verification of the DNA sequence, plasmids were transformed into electrocompetent *L. lactis* using standard techniques[55].

BlaR1 protein expression

BlaR1 constructs were expressed in *L. lactis* PNZ9000 using the nisin-controlled gene expression system[25]. Expression strain seed culture was grown in Difco M17 broth (BD)[56] supplemented with 10 μg ml^{-1} chloramphenicol and 10 g l^{-1} glucose. Large-scale protein expression was performed in a fermenter (Applicon) at 30 °C in M17 BlaR1 Expression Media (M17BEM) and inoculated with seed culture in a 1:100 (v/v) ratio. M17BEM was developed for BlaR1 protein expression based on M17 medium[56] (Supplementary Table 2). After sterilization, the medium was supplemented with 10 μg ml^{-1} chloramphenicol and 100 μM zinc sulfate. The medium was maintained at pH 7 during fermentation with the addition of ammonium hydroxide. Cells were grown to an optical density at 600 nm of around 1.5 before nisin was added to a final concentration of 5 ng ml^{-1}. Protein was expressed for 2–3 h, before collection by centrifugation, then the cells were frozen in liquid nitrogen and stored at −80 °C.

BlaR1 purification

Frozen cell pellets were resuspended in lysis buffer (150 mM sodium chloride; 50 mM HEPES, pH 7.5) with 5 mg ml^{-1} lysozyme (Bio Basic) and 20 µg ml^{-1} DNase I (Roche) and incubated at 37 °C for 1 h with stirring. The cells were then cooled in an ice bath before being lysed with a single pass using a French press (Thermo Fisher Scientific) at ~25,000 psi. All of the subsequent steps were carried out at 4 °C. The lysate was centrifuged at 18,600*g* for 30 min. The supernatant was collected and centrifuged at 40,000 rpm in a Type 45 Ti rotor (Beckman Coulter) for 1 h to pellet the membranes. The membranes were resuspended in lysis buffer using a glass Teflon Dounce homogenizer before being frozen in liquid nitrogen and stored at −80 °C until needed.

Membranes were thawed and resuspended in lysis buffer with 1% (w/v) *n*-dodecyl-β-d-maltopyranoside (DDM) for 1 h to extract the membranes. All detergents used here were acquired from Anatrace. The extracted membranes were centrifuged at 40,000 rpm in a Type 45 Ti rotor (Beckman Coulter) for 45 min to pellet insolubilized membranes. The supernatant was then filtered through a 0.45 µm membrane and imidazole at pH 7.5 was added to 20 mM before being loaded onto a 1 ml HisTrap HP (GE healthcare) equilibrated in buffer A (150 mM sodium chloride; 20 mM HEPES, pH 7.5; 0.016% (w/v) Anagrade DDM or 0.01% (w/v) lauryl maltose neopentyl glycol (LMNG)) supplemented with 20 mM imidazole at 0.5 ml min^{-1}. The column was then sequentially washed with 15 ml buffer A with 20 mM imidazole, 15 ml buffer A with 100 mM imidazole, and eluted in buffer A with 325 mM imidazole. The protein was then immediately desalted into buffer A with a XK-26 column (Amersham Biosciences) packed with Sephadex G-25 medium resin (GE healthcare). Protein was then concentrated using a stirred-cell concentrator with a 100 kDa molecular mass cut-off cellulose membrane (Merck Millipore) before continued purification or flash-freezing in liquid nitrogen followed by storage at −80 °C. BlaR1 concentration was determined by absorbance at 280 nm using the theoretical extinction coefficient, 106,245 M^{-1} cm^{-1}, calculated using ExPASy[57].

During purification of BlaR1, a band directly below the main full-length band was observed on the SDS–PAGE gel (Extended Data Fig. 2e). Edman N-terminal sequencing revealed cleavage was occurring between BlaR1

residues Ser72 and Lys73, corresponding to an exposed region of EL1 and loss of transmembrane helices 1 and 2. As this minor cleavage product has not to our knowledge been observed in BlaR1 expressed in *S. aureus* we expect that this cleavage is due to proteases originating in the *L. lactis* expression system used. We did not see any evidence of this cleavage product in our cryo-EM subclasses or reconstructions.

For cryo-EM, BlaR1 samples in LMNG were further purified using the GraDeR[58] rate zonal ultracentrifugation method. Specifically, 5–25% (v/v) glycerol gradients (light buffer: 5% (v/v) glycerol, 150 mM sodium chloride, 20 mM HEPES, pH 7.5 with or without 0.01% LMNG detergent; heavy buffer: 25% (v/v) glycerol, 150 mM sodium chloride, 20 mM HEPES, pH 7.5) were made using a Gradient Master (BioComp Instruments)[59]. For samples with ampicillin, BlaR1(F284A) was incubated for around 15 min on ice with 1 mM ampicillin and then all subsequent buffers were made with 250 µM ampicillin. After overlaying 150–250 µl of protein sample onto the glycerol gradients, they were immediately centrifuged for the equivalent of 8 h at 55,000 rpm in a SW 55 Ti rotor (Beckman Coulter) at 4 °C. The gradients were then fractionated using a gradient fractionator (BioComp Instruments). The relevant fractions were concentrated using an Amicon Ultra 100 kDa molecular mass cut-off centrifugal filter (Merck Millipore) at 6,000*g* and 4 °C, before being desalted using a Micro Bio-Spin P-30 Chromatography Column (Bio-Rad Laboratories) into buffer B (150 mM sodium chloride and 20 mM HEPES, pH 7.5). The purity was checked on SDS–PAGE gels. For some gels, BlaR1 was incubated with a final concentration of 50 µM BOCILLIN FL[27] for approximately 5 min at room temperature before the addition of loading dye to fluorescently label the sensor domain.

Edman N-terminal sequencing

Samples for Edman N-terminal sequencing were separated on an SDS–PAGE gel before being transferred to a polyvinylidene fluoride membrane and stained using Coomassie. Edman N-terminal sequencing was performed at the SPARK BioCentre.

BlaR1 activity assay with BlaI substrate

Purification of T7-tagged BlaI

The full-length *blaI* gene (UniProt: P0A042) from *S. aureus* was cloned into pET21a with an N-terminal T7-tag using restriction-free cloning[54]. This plasmid was transformed into electrocompetent *E. coli* Rosetta (DE3) cells. Transformed cells were grown at 37 °C in LB medium supplemented with 100 μg ml^{-1} ampicillin. At an OD_{600} of around 0.6–0.8 protein expression was induced by the addition of isopropyl β-d-1-thiogalactopyranoside (IPTG) to 1 mM and the temperature was lowered to 20 °C. Cells were allowed to express protein overnight before collection, flash-freezing in liquid nitrogen and storage at −80 °C.

Protein purification was performed at 4 °C with adaptations from a previous protocol[60]. Cells were resuspended in buffer 2A (50 mM HEPES, pH 7.0) with 1 cOmplete EDTA-free protease inhibitor cocktail tablet (Roche) and DNase I to ~20 μg ml^{-1} and lysed in a French press at 16,000–25,000 psi. BlaI was then purified by ammonium sulfate precipitation with sequential addition of ammonium sulfate to 50%, 60%, 70% and 80% saturation. After each addition of ammonium sulfate, the lysate was incubated for 15 min with inversion followed by centrifugation to pellet the precipitated protein. Protein pellets from 70% and 80% saturated ammonium sulfate solutions were resuspended in buffer 2A and dialysed overnight in buffer 2B (50 mM HEPES, pH 7.0; 5% glycerol). The protein was then loaded on a 5 ml Heparin column (GE Healthcare) equilibrated in buffer 2B at 1 ml min^{-1}. The column was washed with buffer 2B until the absorbance at 280 nm stabilized. Protein was eluted from the column with a linear gradient of buffer 2B to buffer 2C (50 mM HEPES, pH 7.0; 5% glycerol; 2 M sodium chloride) over 60 min at a flow rate of 2 ml min^{-1}. Fractions of interest were concentrated on an Amicon Ultra 10 kDa molecular mass cut-off centrifugal filter (Merck Millipore) and loaded onto a Superdex 75 10/300 column (GE healthcare) equilibrated in buffer 2D (20 mM HEPES, pH 7.0; 500 mM sodium chloride; 5% glycerol). Purified protein was concentrated as before to 4 mg ml^{-1} (as determined by absorbance at 280 nm and correction using the theoretical extinction coefficient 19,940 M^{-1} cm^{-1} calculated using the ExPASy server)[57] before being frozen in liquid nitrogen and stored at −80 °C until needed.

Assaying BlaR1 activity with purified, T7-tagged BlaI

BlaR1(WT), BlaR1(F284A) or BlaR1(USA300) was incubated with BlaI for 4 h at room temperature in assay buffer (150 mM sodium chloride, 20 mM HEPES, pH 7.5, 0.016% DDM). After the incubation period, the reaction was quenched with SDS loading buffer and the samples were analysed using western blotting. The western blot was probed with an anti-T7-tag antibody HRP conjugate (Novagen) diluted 1:15,000 in TTBS (20 mM Tris, 150 mM NaCl, 0.1 % Tween-20) and 3% non-fat skimmed milk powder) for 1 h at room temperature. The blot was developed using Western Chemiluminescent HRP Substrate (Immobilon) and imaged using the Chemi Genius 2 Bio Imaging System (Syngene).

Determining the BlaI cleavage site

To aid Edman N-terminal sequencing, a BlaI–maltose-binding protein (MBP) fusion construct was cloned to the C terminus of BlaI with a Gly-Ser-Ser linker. Initial attempts to transfer the 3 kDa fragment onto the polyvinylidene fluoride membrane failed, but when fused to MBP, the transfer was trivial. The BlaI-MBP construct was designed with an N-terminal decahistidine tag and was cloned into a pGEX-6P-1 expression plasmid with restriction-free cloning[54].

Expression was performed as described for the T7-tagged BlaI construct and cells were lysed as before in buffer 3A (25 mM Tris, pH 8, and 100 mM sodium chloride) with 1 cOmplete EDTA-free protease inhibitor cocktail tablet (Roche) and DNase I (Roche) to ~20 µg ml^{-1}. The cell lysate was centrifuged at 40,000 rpm in a type 45 Ti rotor (Beckman Coulter) for 45 min at 4 °C. The supernatant was filtered through a 0.45 µm filter and imidazole was added to 10 mM before the sample was loaded onto a 1 ml HisTrap HP (GE Healthcare) equilibrated in buffer 3A at 0.75 ml min^{-1}. The column was then washed with 10 cv of 2% buffer 3B (50 mM Tris pH 8, 500 mM imidazole and 100 mM sodium chloride), before being eluted on a gradient of 0–100% over 40 min at flow rate of 1 ml min^{-1}. Fractions containing the protein of interest were concentrated on an Amicon Ultra 30 kDa molecular mass cut-off centrifugal filter (Merck Millipore). The concentrated protein was loaded onto the Superdex 75 10/300 column (GE

healthcare) equilibrated in buffer 3C (25 mM Tris, pH 8.0, 100 mM sodium chloride). Purified protein was concentrated as before, frozen in liquid nitrogen and stored at −80 °C until needed.

Inductively coupled plasma mass spectrometry

BlaR1(F284A) was purified in LMNG using immobilized metal affinity chromatography before being concentrated and purified by SEC in buffer B. Protein and buffer samples were prepared in 1% HNO_3 with ^{45}Sc (100 ppb) as an internal standard. Inductively coupled plasma mass spectrometry to determine zinc concentration was performed using the NexION 300D (Perkin Elmer) system equipped with a SC-2 DX autosampler, DXi-FAST micro-peristaltic pump, a cyclonic spray chamber, a triple cone interface, a quadrupole ion deflector and Universal Cell Technology. Calibration was performed using IV-Stock-4 calibration standard (Inorganic Ventures). All elements were run in reaction mode (using Dynamic Reaction Cell technology) using ammonia as a reaction gas to remove potential polyatomic interferences. The detection limit for ^{66}Zn was determined to be 0.765 ppb.

Cryo-electron microscopy of BlaR1

For all cryo-EM datasets, C-flat (1.2/1.3) 300 mesh grids (Protochips) were glow-discharged before 3 µl of protein sample at an absorbance at 280 nm of 0.27–0.35 was applied to the grid. The grids were blotted for 1–2 s using the Vitrobot Mark IV (Thermo Fisher Scientific) at 4 °C and 100% humidity before being plunged into liquid ethane. The grids were transferred to liquid nitrogen for storage until imaging.

Cryo-EM grids were screened at the High-resolution Macromolecular Cryo-Electron Microscopy facility at the University of British Columbia on a Glacios (Thermo Fisher Scientific) transmission cryogenic electron microscope with a 200 kV accelerating voltage and a Falcon 3 camera using the EPU (Thermo Fisher Scientific) software package.

The BlaR1(WT) dataset was collected at the High-resolution Macromolecular Cryo-Electron Microscopy facility at the University of British Columbia using the 300 kV Titan Krios (Thermo Fisher Scientific)

transmission cryogenic electron microscope equipped with a Falcon 4i (Thermo Fisher Scientific) direct electron detector and a Selectris (Thermo Fisher Scientific) energy filter. Images were gathered in electron-counting mode with a calibrated magnification of ×165,000, corresponding to 0.77 Å per physical pixel. The total dose delivered to the grids was 50 e^{-} $Å^{-2}$ over 875 frames in the EER video format. Fully automated data collection was performed using EPU (Thermo Fisher Scientific) with a nominal defocus range set from 0.5 to 3.0 μm.

The BlaR1(F284A) ligand-free dataset was collected at the Pacific Northwest Cryo-EM Centre using a 300 kV Titan Krios (Thermo Fisher Scientific) transmission cryogenic electron microscope equipped with the GIF Quantum energy filter (Gatan; operated at 20 keV) and a K3 Summit (Gatan) direct electron detector. Images were gathered in super-resolution mode at a calibrated magnification of 81,000x, corresponding to 1.059 Å per physical pixel (0.5295 Å per super-resolution pixel). A total dose of 33 e^{-} $Å^{-2}$ was delivered to the grids over 51 frames. Fully automated data collection was performed using SerialEM with a nominal defocus range set from 1.5 to 3.5 μm (ref. [61]).

The BlaR1(F284A)–ampicillin dataset was collected at the HHMI Janelia Research Campus using the 300 kV Titan Krios (Thermo Fisher Scientific) transmission cryogenic electron microscope equipped with a spherical aberration corrector, a GIF Quantum energy filter (Gatan) operated at 20 keV and a K3 Summit (Gatan) direct electron detector. Images were gathered in super-resolution mode at a calibrated magnification of ×64,000, corresponding to 1.065 Å per physical pixel (0.5325 Å per super-resolution pixel). The total dose delivered to the grids was 60 e^{-} $Å^{-2}$ over 60 frames. Fully automated data collection was performed using SerialEM with a nominal defocus range set from 1.0 to 2.0 μm (ref. [61]).

Cryo-EM data processing

For the BlaR1(WT) sample, 16,599 videos were collected at a pixel size of 0.77 Å per pixel. Motion correction was performed using MotionCor2[62] and the contrast transfer functions (CTFs) of the summed and dose-weighted micrographs were determined using CTFFIND4[63]. Approximately 2,000

particles were manually selected to generate reference-free 2D-class averages in RELION[64]. Representative 2D classes were used as templates for automated particle picking in RELION (v.3.1 and v.4) with 7,508,828 particles picked. 2D classification was performed using CryoSPARC (v.3.2 and v.3.3.2)[65], resulting in 4,286,427 particles taken forward for further processing. Three initial volumes were generated and processed for multiple rounds of heterogeneous refinement. The best volume was further refined with NU-refinement with C_2 symmetry imposed and used as the initial model for additional rounds of heterogenous refinement to recover good particles that were previously discarded in the 2D classification steps. A particle dataset of 286,692 particles was transferred back to RELION with cs2star in the pyem[66] software package and processed for multiple rounds of 3D classification. The particles associated with the best volumes were grouped, refined and used for Bayesian polishing[67] and CTF refinement[68]. Partial signal subtraction was used to remove the micelle density. A reconstruction with C_2 symmetry applied was generated from 143,227 particles to a resolution of 4.2 Å as determined using the gold-standard refinement procedure with an FSC of 0.143 and high-resolution noise substitution to correct soft-mask effects[69]. Symmetry expansion followed by further 3D classification without alignment using a mask covering the sensor domains identified different sensor domain states. A representative class with the most resolved EL1 region contained 23,797 unique particles and refined to 4.9 Å resolution with no symmetry applied (C_1). Signal subtraction to remove the sensor domain density was carried out on the 143,227 particles again followed by symmetry expansion and 3D classification without alignment resulting in 66,347 unique particles that refined to 3.8 Å resolution with no symmetry applied (C_1) (Extended Data Table 1 and Extended Data Figs. 3 and 4).

For the BlaR1(F284A) sample, a total of 11,844 video series was collected in super-resolution mode at 0.5295 Å per pixel and binned to 1.059 Å per pixel during motion correction performed with MotionCor2[62]. The CTFs of the summed and dose-weighted micrographs were determined using CTFFIND4[63]. Particle picking was carried out as described above with 4,181,855 particles automatically picked. Reference-free 2D classification was performed in CryoSPARC[65], resulting in 534,610 good particles taken

for further processing. Further particle sorting was performed as described above and 221,007 particles were transferred back to RELION and processed for multiple rounds of 3D classification, and the particles associated with the best volumes were grouped, refined, and used for Bayesian polishing[67] and CTF refinement[68], and the micelle density was removed through partial signal subtraction. One round of 3D classification without alignment was performed and the particles belonging to the best volumes were grouped and refined. The resulting volume after postprocessing reached a resolution of 4.3 Å with C_2 symmetry averaging applied (Extended Data Fig. 3). An example data-processing overview is included in Extended Data Fig. 5.

For the BlaR1(F284A)–ampicillin sample, 10,872 video series were collected in super-resolution mode at 0.5325 Å per pixel and binned to 1.065 Å per pixel during motion correction performed using MotionCor2[62]. The CTFs of the summed and dose-weighted micrographs were determined using CTFFIND4[63]. Particle picking was performed as described above with 5,613,933 particles automatically picked. Reference-free 2D classification was performed in CryoSPARC[65], resulting in 216,832 good particles. Further particle sorting was carried out using heterogeneous refinement as described above and 245,700 particles were transferred back to RELION. Multiple rounds of 3D classification were carried out with the particles associated with the best volumes grouped, refined, used for Bayesian polishing[67] and CTF refinement[68], and the micelle density was removed through partial signal subtraction. One round of 3D classification without alignment was performed and the particles belonging to the best volumes were grouped and refined. The resulting volume after post-processing reached a resolution of 4.6 Å with C_2 symmetry averaging applied (Extended Data Fig. 3).

Model building and refinement

A prediction of the transmembrane and catalytic regions of the BlaR1 structure from HHPred[70] and Modeller[71] was used as a starting point for model building using the BlaR1(WT) transmembrane reconstruction to 3.8 Å resolution. Incorrect regions were deleted and iterative model building and

refinement carried out in Coot[72], Phenix real.space refine[73] in the Phenix software package[74], and density-guided Rosetta refinement with symmetry[39]. For the sensor domain, the X-ray crystal structure of the *S. aureus* sensor domain (PDB: 1XA7) was automatically fit into the 4.2 Å C_2 reconstruction using Phenix dock_in_map. Although this region was of lower local resolution (Extended Data Fig. 3), clear secondary structure features enabled confident placement (Extended Data Fig. 7a–c). Residues 52–96, forming the extracellular facing loop EL1, were not clearly resolved in entirety. AlphaFold2 (not available at the time of initial model building) reproducibly and confidently predicted the C-terminal region of this loop as bound to the sensor-domain β-lactam binding groove, which was supported by low-resolution features in the reconstruction (Extended Data Fig. 6g–i). The EL1 loop conformation for residues 87–97 from AlphaFold2 was incorporated into the final model and subject to extensive density restrained refinement in Rosetta and Phenix real.space refine. Sensor-domain loop 413–427 was omitted due to interchain clashes. For the BlaR1(WT) C_1 (4.9 Å) reconstruction with shifted sensor-domain positions, residues 413–427 did not clash and were included. For the BlaR1(F284A) (4.3 Å) and BlaR1(F284A)–ampicillin (4.6 Å) reconstructions, the transmembrane/catalytic regions and sensor domains were docked into the density. The relative distance between sensor and transmembrane domains was different between BlaR1(WT) (furthest), BlaR1(F284A) and BlaR1(F284A)–ampicillin (closest) with the loop (residues 332–345) joining the two adopting different relative positions. For BlaR1(F284A), the final model consists of residues 1–55 and 87–585 and, for BlaR1(F284A)–ampicillin, residues 1–55, 97–412 and 428–585 (EL1 residues 87–96 and sensor-domain residues 413–427 were omitted). Two active site zincs were included in both cases (Extended Data Table 1 and Extended Data Figs. 3, 4 and 6).

Model validation was performed using MolProbity as implemented in Phenix[75,76]. Figures were created using PyMol (Schrödinger), UCSF Chimera[77], UCSF ChimeraX[78] and Adobe Illustrator 2021 (Adobe). BlaR1 was inserted into a membrane generated with the CHARM-GUI membrane builder[79]. Protein interface analysis was carried out with the PISA server using the default parameters[80]. Sequence conservation was analysed using the ConSurf server[81].

BlaR1 functional assays carried out in *S. aureus*

The *S. aureus* strains used in the assay (Supplementary Table 3) were cultured in tryptic soy broth (TSB) at 37 °C at 180 rpm. The natural plasmid of SF8300 (a USA300 strain) that contains the *blaR1-blaI* regulatory genes (pUSA300HOUMR) was used as a template for PCR amplification. Different variants of the *blaR1-blaI* constructs were generated by splice overlap PCR and cloned into a constitutively expressing vector, pTX_{Δ}[82]. The plasmids were then transformed to RN4220 through electroporation. After sequence verification, the plasmids were introduced to SF8300ermS (erythromycin sensitive SF8300) through transduction using the phage, Φ11. SF8300ermS was obtained by evicting pUSA300HOUMR through serial passaging of SF8300 in TSB and by selecting for an erythromycin-sensitive colony[83]. pUSA300HOUMR, in addition to containing *blaZ* and its control elements (*blaR1-blaI*), also contains an erythromycin-resistance gene[84]. Thus, the SF8300ermS lacked β-lactamase and its control elements (*blaZ-blaR1-blaI*). Lists of the plasmids and primers used in this assay are provided in Supplementary Tables 4 and 5, respectively. Strains containing the pTX_{Δ} plasmid were grown and maintained in 12.5 μg ml^{-1} tetracycline. *blaR1* and *blaI* are expressed from a plasmid and in turn control the expression of chromosomally encoded *mecA*. For this purpose, *blaR1(F284A)-blaI* was cloned into a constitutively expressed plasmid and transduced into SF8300ermS that lacked the natural plasmid (pUSA300HOUMR) containing *blaZ* and its regulatory elements (*blaR1-blaI)*. SF8300ermS transduced with an identical plasmid with either an empty vector, a fully functional *blaR1-blaI* or a functionally attenuated *blaR1(null)-blaI* and *blaR1(null)-blaI(null)* constructs were used as controls. The null mutants were created by mutating the methionine start codon of *blaR1* and *blaI* to threonine, leading to strains deficient of these proteins. The isogenic strains thus created were assessed for their growth properties either in absence or in presence of nafcillin, an easily available and prototype β-lactam antibiotic that is structurally similar to methicillin.

Growth assays were performed using the Bioscreen C Pro system (Growth Curves USA). The initial bacterial OD_{600} was adjusted to 0.1 in TSB containing 2 μg ml^{-1} nafcillin or no nafcillin and 200 μl of this mixture was

added to a Honeycomb microplate (Growth Curves USA) in triplicates. Wells containing TSB without any antibiotics were included as controls. The readings were recorded every 30 min for 24 h at 600 nm. The temperature was set to 37 °C. The settling time was set to 30 s and the shake duration was 15 s. DNA sequence analysis was performed using DNAstar software. All experiments except for sequencing were repeated at least twice. The fidelity of all the plasmid constructs and point mutations described in this study were validated through Sanger sequencing.

BlaR1 cleavage with *S. aureus* signal peptidase SpsB

The gene for *S. aureus* MRSA252 SpsB (UniProt: Q6GIC3) was cloned into the expression vector pET28a+ (Novagen) using the NdeI and XhoI sites. The expressed protein has an N-terminal hexahistidine tag, followed by a thrombin cleavage site (MGSSHHHHHHSSGLVPRGSHM). Plasmid was transformed into *E. coli* BL21(DE3). One litre of LB medium supplemented with kanamycin (0.05 mg ml^{-1}) was inoculated with 50 ml of overnight culture. Cells were grown at 37 °C with shaking (250 rpm) until an OD_{600} of 0.6, at which point 1 mM of IPTG was added to the culture to induce protein expression. Cells were grown for 4 h after induction, then pelleted by centrifugation (6,000*g*, for 5 min at 4 °C). Then, 5 g of the cell pellet was resuspended in 25 ml of buffer A (20 mM Tris pH 8.0, 100 mM NaCl) and then sonicated using a Sonic Dismembrator Model 500 (Thermo Fisher Scientific), with a 15 s pulse and 30 s rest cycle three times at 30% amplitude, on ice and lysed by passing through an EmulsiFlex-C3 cell disruptor (Avestin) for 5 min. The resulting cell lysate was clarified by centrifugation (35,000*g* for 30 min at 4 °C). The pellet was resuspended in 25 ml of buffer A with 0.1% Triton X-100 and dounced 50 times followed by centrifugation (35,000*g* for 30 min at 4 °C). The clarified supernatant was then subjected to Ni^{2+}-NTA affinity chromatography purification. The clarified supernatant, containing the His-tagged SpsB, was poured onto a column containing 3 ml of Ni^{2+}-charged resin (Qiagen) pre-equilibrated with kinetic buffer (buffer A with 0.01% DDM). The column was washed twice with 30 ml of kinetic buffer containing first 10 mM, and then 20 mM imidazole. Next, His-tagged SpsB was eluted using a step gradient, created by 6 ml of kinetic buffer containing imidazole in increasing concentrations of 100 to 500 mM imidazole. Fractions containing purified SpsB were

concentrated to 1 ml using an Amicon ultracentrifugal filter (Millipore) with a 3 kDa cut-off and dialysed against 1 l of kinetic buffer at 4 °C overnight. Protein concentration was determined by BCA assay (Thermo Fisher Scientific) according to the manufacturer's protocol.

To assay for SpsB-mediated BlaR1 cleavage, BlaR1(WT) or BlaR1(G331K) in reaction buffer (150 mM sodium chloride; 20 mM HEPES, pH 7.5, 0.01% (w/v) LMNG) was incubated overnight at room temperature with an excess molar ratio of SpsB. The resulting sample was incubated with BOCILLIN-FL[27] for 5 min before SDS–PAGE.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Models and cryo-EM reconstructions for BlaR1(WT), BlaR1(F284A) ampicillin-free and BlaR1(F284A) ampicillin-bound samples have been deposited at the PDB and the EMDB under accession codes 8EXP, 8EXQ, 8EXR, 8EXS and 8EXT, and EMD-28658, EMD-28659, EMD-28660, EMD-28661 and EMD-28662, respectively.

References

1. Turner, N. A. et al. Methicillin-resistant *Staphylococcus aureus*: an overview of basic and clinical research. *Nat. Rev. Microbiol.* **17**, 203–218 (2019).

2. Murray, C. J. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).

3. Hackbarth, C. J. & Chambers, H. F. *blaI* and *blaR1* regulate β-lactamase and PBP 2a production in methicillin-resistant *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **37**, 1144–1149 (1993).

4. Zhang, H. Z., Hackbarth, C. J., Chansky, K. M. & Chambers, H. F. A proteolytic transmembrane signaling pathway and resistance to β-lactams in staphylococci. *Science* **291**, 1962–1965 (2001).

5. Herzberg, O. & Moult, J. Bacterial resistance to β-lactam antibiotics: crystal structure of β-lactamase from *Staphylococcus aureus* PC1 at 2.5 Å resolution. *Science* **236**, 694–701 (1987).

6. Lim, D. & Strynadka, N. C. J. Structural basis for the β-lactam resistance of PBP2a from methicillin-resistant *Staphylococcus aureus*. *Nat. Struct. Biol.* **9**, 870–876 (2002).

7. Blázquez, B. et al. Regulation of the expression of the β-lactam antibiotic-resistance determinants in methicillin-resistant *Staphylococcus aureus* (MRSA). *Biochemistry* **53**, 1548–1550 (2014).

8. Amoroso, A. et al. A peptidoglycan fragment triggers β-lactam resistance in *Bacillus licheniformis*. *PLoS Pathog.* **8**, e1002571 (2012).

9. Rodvold, K. A. & Mcconeghy, K. W. Methicillin-resistant *Staphylococcus aureus* therapy: past, present, and future. *Clin. Infect. Dis.* **58**, S20–S27 (2014).

10. Wu, Q. et al. Systematic review and meta-analysis of the epidemiology of vancomycin-resistance *Staphylococcus aureus* isolates. *Antimicrob. Resist. Infect. Control* **10**, 101 (2021).

11. Rossi, F. et al. Transferable vancomycin resistance in a community-associated MRSA lineage. *N. Engl. J. Med.* **370**, 1524–1531 (2014).

12. Otero, L. H. et al. How allosteric control of *Staphylococcus aureus* penicillin binding protein 2a enables methicillin resistance and physiological function. *Proc. Natl Acad. Sci. USA* **110**, 16808–16813 (2013).

13. Katayama, Y., Ito, T. & Hiramatsu, K. A new class of genetic element, *Staphylococcus* cassette chromosome *mec*, encodes methicillin resistance in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **44**, 1549–1555 (2000).

14. Arêde, P., Ministro, J. & Oliveira, D. C. Redefining the role of the β-lactamase locus in methicillin-resistant *Staphylococcus aureus*: β-lactamase regulators disrupt the MecI-mediated strong repression on mecA and optimize the phenotypic expression of resistance in strains with constitutive mecA. *Antimicrob. Agents Chemother.* **57**, 3037–3045 (2013).

15. Alexander, J. A. N. et al. Structural analysis of avibactam-mediated activation of the bla and mec divergons in methicillin-resistant *Staphylococcus aureus*. *J. Biol. Chem.* **295**, 10870–10884 (2020).

16. McKinney, T. K., Sharma, V. K., Craig, W. A. & Archer, G. L. Transcription of the gene mediating methicillin resistance in *Staphylococcus aureus* (*mecA*) is corepressed but not coinduced by cognate *mecA* and β-lactamase regulators transcription of the gene mediating methicillin resistance in *Staphylococcus aureus*. *J. Bacteriol.* **183**, 6862–6868 (2001).

17. Lakhundi, S. & Zhang, K. Methicillin-resistant *Staphylococcus aureus*: molecular characterization, evolution, and epidemiology. *Clin Microbiol. Rev.* **31**, e00020-18 (2018).

18. Liu, C. et al. A population-based study of the incidence and molecular epidemiology of methicillin-resistant *Staphylococcus aureus* disease in San Francisco, 2004–2005. *Clin. Infect. Dis.* **46**, 1637–1646 (2008).

19. Hanique, S. et al. Evidence of an intramolecular interaction between the two domains of the BlaR1 penicillin receptor during the signal transduction. *J. Biol. Chem.* **279**, 14264–14272 (2004).

20. Frederick, T. E., Wilson, B. D., Cha, J., Mobashery, S. & Peng, J. W. Revealing cell-surface intramolecular interactions in the BlaR1 protein of methicillin-resistant *Staphylococcus aureus* by NMR spectroscopy. *Biochemistry* **53**, 10–12 (2014).

21. Belluzo, B. S. et al. An experiment-informed signal transduction model for the role of the *Staphylococcus aureus* MecR1 protein in β-lactam resistance. *Sci. Rep.* **9**, 19558 (2019).

22. Sala, C. et al. Genome-wide regulon and crystal structure of BlaI (Rv1846c) from *Mycobacterium tuberculosis*. *Mol. Microbiol.* **71**, 1102–1116 (2009).

23. Sandhu, B. K., Edwards, A. N., Anderson, S. E., Woods, E. C. & McBride, S. M. Regulation and anaerobic function of the *Clostridioides difficile* β-lactamase. *Antimicrob. Agents Chemother.* **64**, e01496-19 (2019).

24. Sobhanifar, S., Prehna, G. & Strynadka, N. in *Handbook of Proteolytic Enzymes* (eds Barrett, A. et al.) 1237–1243 (Academic, 2012).

25. Mierau, I. & Kleerebezem, M. 10 years of the nisin-controlled gene expression system (NICE) in *Lactococcus lactis*. *Appl. Microbiol. Biotechnol.* **68**, 705–717 (2005).

26. Llarrull, L. I. & Mobashery, S. Dissection of events in the resistance to β-lactam antibiotics mediated by the protein BlaR1 from *Staphylococcus aureus*. *Biochemistry* **51**, 4642–4649 (2012).

27. Zhao, G., Meier, T. I., Kahl, S. D., Gee, K. R. & Blaszczak, L. C. BOCILLIN FL, a sensitive and commercially available reagent for detection of penicillin-binding proteins. *Antimicrob. Agents Chemother.* **43**, 1124–1128 (1999).

28. Llarrull, L. I., Toth, M., Champion, M. M. & Mobashery, S. Activation of BlaR1 protein of methicillin-resistant *Staphylococcus aureus*, its proteolytic processing, and recovery from induction of resistance. *J. Biol. Chem.* **286**, 38148–38158 (2011).

29. Viklund, H., Granseth, E. & Elofsson, A. Structural classification and prediction of reentrant regions in α-helical transmembrane proteins: application to complete genomes. *J. Mol. Biol.* **361**, 591–603 (2006).

30. Jongeneel, C. V., Bouvier, J. & Bairoch, A. A unique signature identifies a family of zinc-dependent metallopeptidases. *FEBS Lett.* **242**, 211–214 (1989).

31. Granseth, E., Daley, D. O., Rapp, M., Melén, K. & von Heijne, G. Experimentally constrained topology models for 51,208 bacterial inner membrane proteins. *J. Mol. Biol.* **352**, 489–494 (2005).

32. Daley, D. O. et al. Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science* **308**, 1321–1323 (2005).

33. White, D. C. & Frerman, F. E. Extraction, characterization, and cellular localization of the lipids of *Staphylococcus aureus*. *J. Bacteriol.* **94**, 1854–1867 (1967).

34. Young, S. A., Desbois, A. P. & Coote, P. J. Characterisation of *Staphylococcus aureus* lipids by nanoelectrospray ionisation tandem mass spectrometry (nESI-MS/MS). Preprint at *bioRxiv* https://doi.org/10.1101/593483 (2019).

35. Schleifer, K. H. et al. Transfer of *Streptococcus lactis* and related Streptococci to the genus *Lactococcus* gen. nov. *Syst. Appl. Microbiol.* **6**, 183–195 (1985).

36. Driessen, A. J. M., Zheng, T., Veld, G. I. T., Op Den Kamp, J. A. F. & Konings, W. N. Lipid requirement of the branched-chain amino acid transport system of *Streptococcus cremoris*. *Biochemistry* **27**, 865–872 (1988).

37. Wilke, M. S., Hills, T. L., Zhang, H.-Z., Chambers, H. F. & Strynadka, N. C. J. Crystal structures of the Apo and penicillin-acylated forms of the BlaR1 β-lactam sensor of *Staphylococcus aureus*. *J. Biol. Chem.* **279**, 47278–47287 (2004).

38. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

39. Dimaio, F. et al. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* **12**, 361–365 (2015).

40. Staude, M. W. et al. Investigation of signal transduction routes within the sensor/transducer protein BlaR1 of *Staphylococcus aureus*.

Biochemistry **54**, 1600–1610 (2015).

41. Hooper, N. M. Families of zinc metalloproteases. *FEBS Lett.* **354**, 1–6 (1994).

42. Marie-Claire, C. et al. Evidence by site-directed mutagenesis that arginine 203 of thermolysin and arginine 717 of neprilysin (neutral endopeptidase) play equivalent critical roles in substrate hydrolysis and inhibitor binding. *Biochemistry* **36**, 13938–13945 (1997).

43. Tyndall, J. D. A., Nall, T. & Fairlie, D. P. Proteases universally recognize beta strands in their active sites. *Chem. Rev.* **105**, 973–999 (2005).

44. Gomis-Rüth, F. X., Botelho, T. O. & Bode, W. A standard orientation for metallopeptidases. *Biochim. Biophys. Acta* **1824**, 157–163 (2012).

45. Punjani, A. & Fleet, D. J. 3D variability analysis: resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702 (2021).

46. Kobayashi, T., Zhu, Y. F., Nicholls, N. J. & Lampen, J. O. A second regulatory gene, *blaR1*, encoding a potential penicillin-binding protein required for induction of β-lactamase in *Bacillus licheniformis*. *J. Bacteriol.* **169**, 3873–3878 (1987).

47. Quigley, A. et al. The structural basis of ZMPSTE24-dependent laminopathies. *Science* **339**, 1604–1607 (2013).

48. Pryor, E. E. et al. Structure of the integral membrane protein CAAX protease Ste24p. *Science* **339**, 1600–1604 (2013).

49. Berzigotti, S., Benlafya, K., Sépulchre, J., Amoroso, A. & Joris, B. *Bacillus licheniformis* BlaR1 L3 loop is a zinc metalloprotease activated by self-proteolysis. *PLoS ONE* **7**, e36400 (2012).

50. Beaumont, A. et al. The role of histidine 231 in thermolysin-like enzymes: a site-directed mutagenesis study. *J. Biol. Chem.* **270**, 16803–16808 (1995).

51. Boudreau, M. A., Fishovitz, J., Llarrull, L. I., Xiao, Q. & Mobashery, S. Phosphorylation of BlaR1 in manifestation of antibiotic resistance in methicillin-resistant *Staphylococcus aureus* and its abrogation by small molecules. *ACS Infect. Dis.* **1**, 454–459 (2015).

52. Powers, M. E. et al. Type I signal peptidase and protein secretion in *Staphylococcus epidermidis*. *J. Bacteriol.* **193**, 2677–2686 (2011).

53. Kuipers, O. P., de Ruyter, P. G. G. A., Kleerebezem, M. & de Vos, W. M. Quorum sensing-controlled gene expression in lactic acid bacteria. *J. Biotechnol.* **64**, 15–21 (1998).

54. van den Ent, F. & Löwe, J. RF cloning: a restriction-free method for inserting target genes into plasmids. *J. Biochem. Biophys. Methods* **67**, 67–74 (2006).

55. Geertsma, E. R. & Poolman, B. High-throughput cloning and expression in recalcitrant bacteria. *Nat. Methods* **4**, 705–707 (2007).

56. Terzaghi, B. & Sandine, W. Improved medium for lactic streptococci and their bacteriophages. *Appl. Microbiol.* **29**, 807–813 (1975).

57. Gasteiger, E. et al. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Springer, 2005); https://doi.org/10.1385/1-59259-890-0:571

58. Hauer, F. et al. GraDeR: membrane protein complex preparation for single-particle cryo-EM. *Structure* **23**, 1769–1775 (2015).

59. Coombs, D. H. & Watts, N. R. M. Generating sucrose gradients in three minutes by tilted tube rotation. *Anal. Biochem.* **148**, 254–259 (1985).

60. Grossman, M. J. & Lampen, J. O. Purification and DNA binding properties of the blaI gene product, repressor for the (3-lactamase gene, blaP, of *Bacillus licheniformis*. *Nucleic Acids Res.* **15**, 6049–6062 (1987).

61. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51

(2005).

62. Zheng, S. Q. et al. MotionCor2—anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).

63. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).

64. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).

65. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

66. Asarnow, D., Palovcak, E. & Cheng, Y. UCSF pyem v0.5. *Zenodo* https://doi.org/10.5281/zenodo.3576630 (2019).

67. Zivanov, J., Nakane, T. & Scheres, S. H. W. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCrJ* **6**, 5–17 (2019).

68. Zivanov, J., Nakane, T. & Scheres, S. H. W. Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1. *IUCrJ* **7**, 253–267 (2020).

69. Chen, S. et al. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135**, 24–35 (2013).

70. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).

71. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinform.* **54**, 5.6.1–5.6.37 (2016).

72. Casañal, A., Lohkamp, B. & Emsley, P. Current developments in Coot for macromolecular model building of electron cryo-microscopy and crystallographic data. *Protein Sci.* **29**, 1069–1078 (2020).

73. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).

74. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).

75. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293 (2018).

76. Afonine, P. V. et al. New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr. D* **74**, 814–840 (2018).

77. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

78. Pettersen, E. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

79. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).

80. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

81. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).

82. Chan, L. C. et al. PBP 4 mediates high-level resistance to new-generation cephalosporins in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **60**, 3934–3941 (2016).

83. Chan, L. C. et al. Ceftobiprole- and Ceftaroline-resistant methicillin-resistant *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **59**, 2960–2963 (2015).

84. Lowy, F. D. Antimicrobial resistance: the example of *Staphylococcus aureus*. *J. Clin. Invest.* **111**, 1265–1273 (2003).

Acknowledgements

We thank the Canadian Foundation of Innovation and BC Knowledge Development Fund for infrastructure at the High Resolution Macromolecular EM facility at UBC and the members of the operational team C. Atkinson, F. Rossmann and J. Felt; X. Zhao and Z. Yu for help with earlier cryo-EM data collection at the Howard Hughes Medical Institute CryoEM Shared Resources facility and, more recently, H. Scott for assistance with cryo-EM data collection at the Pacific Northwest Cryo-EM Center (PNCC); and T. Richardson-Sanchez in the laboratory of M. Murphy for running the inductively coupled plasma mass spectrometry samples. A portion of this research was supported by NIH grant U24GM129547 and performed at the PNCC at OHSU and accessed through EMSL (grid.436923.9); a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research. Molecular graphics and analyses were performed using UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases. This work was funded by the Canadian Institute of Health Research (to N.C.J.S.) and the National Institutes of Health (2R01AI100291 and R21AI142501, to S.S.C.). N.C.J.S. is a Canada Research Chair Tier 1. J.A.N.A. thanks the Vanier Graduate Scholarship Program for salary funding during this work.

Author information

Author notes

1. These authors contributed equally: J. Andrew N. Alexander, Liam J. Worrall

Authors and Affiliations

1. Department of Biochemistry and Molecular Biology, The University of British Columbia, Vancouver, British Columbia, Canada

   J. Andrew N. Alexander, Liam J. Worrall, Jinhong Hu, Marija Vuckovic, Solmaz Sobhanifar, Federico I. Rosell, Joshua Jenkins, Wesley A. Mosimann & Natalie C. J. Strynadka

2. Centre for Blood Research, The University of British Columbia, Vancouver, British Columbia, Canada

   J. Andrew N. Alexander, Liam J. Worrall, Jinhong Hu, Marija Vuckovic, Solmaz Sobhanifar, Federico I. Rosell, Joshua Jenkins, Wesley A. Mosimann & Natalie C. J. Strynadka

3. HRMEM Facility, The University of British Columbia, Vancouver, British Columbia, Canada

   Liam J. Worrall & Natalie C. J. Strynadka

4. Department of Microbial Pathogenesis, School of Dentistry, University of Maryland, Baltimore, MD, USA

   Nidhi Satishkumar, Raymond Poon & Som S. Chatterjee

5. Institute of Marine and Environmental Technology, Baltimore, MD, USA

   Nidhi Satishkumar, Raymond Poon & Som S. Chatterjee

6. Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada

   Daniel Chiang & Mark Paetzel

7. Division of Infectious Diseases, School of Medicine, University of California, San Francisco, San Francisco, CA, USA

   Henry F. Chambers

Contributions

J.A.N.A., M.V., D.C., M.P., N.S. and R.P. performed cloning. J.A.N.A., M.V., F.I.R., J.J., D.C. and M.P. performed protein purification. J.H., L.J.W. and J.A.N.A. performed cryo-EM data processing. L.J.W. carried out model building and refinement. L.J.W., J.A.N.A., N.C.J.S. and M.P. performed structure analyses. J.A.N.A., M.V., W.A.M. and S.S. carried out in vitro activity assays. N.S., R.P. and S.S.C. carried out in vivo activity assays. J.A.N.A., L.J.W. and N.C.J.S. wrote the manuscript with input from H.F.C. and S.S. All of the authors read the paper.

Corresponding author

Correspondence to Natalie C. J. Strynadka.

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks John Rubinstein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Amino acid sequence alignment of *S. aureus* BlaR1 homologues: *S. aureus* BlaR1 (UniProt ID: Q00419), BlaR1 from MRSA strain USA300 (Refseq ID: WP_001096381.1), *S. aureus* MecR1 (UniProt ID: P0A0B1), and *Clostridioides difficile* BlaR (UniProt ID: Q188R3).

S. aureus BlaR1 secondary structure is indicated above the sequence alignment.

Extended Data Fig. 2 Characterization of *S. aureus* BlaR1 by glycerol gradient centrifugation, SDS-PAGE and size exclusion chromatography.

a, WT BlaR1 in membranes cleaves T7-tagged BlaI as shown by an anti-T7-tag antibody probed Western blot. Representative gel shown from three independent replicates. **b**, Glycerol gradient chromatogram of BlaR1(WT). **c**, Coomassie stained (left) and fluorescent (BOCILLIN FL) imaged (right) SDS-PAGE gel of BlaR1(WT) fractions from a glycerol gradient. Samples were incubated with BOCILLIN FL, a fluorescently labelled β-lactam antibiotic, prior to SDS-PAGE. BlaR1(WT) undergoes spontaneous autocleavage. BlaR1(WT) full-length (FL), the BlaR1(WT) C-terminal autocleavage product including residues 284-585 (C-CL), the BlaR1(WT) N-terminal autocleavage product including residues 1-283 (N-CL) and BlaR1(WT) sensor domain including residues 332-585 (SD) bands are indicated. Representative gel shown from multiple (>10) independent purifications. **d**, Glycerol gradient chromatogram of BlaR1(F284A) with and without ampicillin. **e**, Coomassie stained (left) and fluorescent (BOCILLIN FL) imaged (right) SDS-PAGE gel of BlaR1(F284A) fractions from the glycerol gradients with and without ampicillin as indicated. Acylation with ampicillin confirmed by inhibition of BOCILLIN binding. Representative gel shown from multiple (>5) independent purifications. **f**, Size exclusion

chromatography using a Superdex 200, 5/150 column shows BlaR1(WT) and BlaR1(F284A) elute as single peaks. The later elution profile of the BlaR1(F284A) sample suggests it may form a more compact dimer. **g**, Purified BlaR1(WT) undergoes spontaneous cleavage. Samples were loaded on an SDS-PAGE gel following incubation for 1 day following purification at the temperature indicated. Representative gel shown from two independent replicates. **h**, Schematic of the BlaR1 cleavage products. Molecular weights of the protein ladder components are indicated in kDa for all gels. Gels in panels C and F were scanned for fluorescence using a Chemi Genius 2 Bio Imaging System (Syngene). For gel source data, see Supplementary Fig. 1.

Extended Data Fig. 3 Single particle cryo-EM data processing of BlaR1.

a, Representative micrographs for BlaR1(WT), BlaR1(F284A) and BlaR1(F284A)-ampicillin, respectively. **b**, Representative 2D classes for BlaR1(WT), BlaR1(F284A) and BlaR1(F284A)-ampicillin, respectively. **c**, 3D reconstructions BlaR1(WT) C2, BlaR1(WT) C1, BlaR1(WT) TM, BlaR1(F284A) and BlaR1(F284A)-ampicillin, respectively. **d**, Gold-standard half-map FSC curves as in C (black line) and model-map FSC curves (grey line). Dashed line at 0.143 FSC. **e**, As **c** coloured according to local resolution.

Extended Data Fig. 4 Example cryo-EM density displayed for the BlaR1(WT) map.

a, Density for BlaR1(WT) model secondary structure elements shown in mesh. **b** and **c**, Stereo views of active site with an intact autocleavage loop (**b**) or cleaved autocleavage loop (**c**).

Extended Data Fig. 5 Example single particle cryo-EM data processing strategy of ligand-free BlaR1.

Blue and green arrows correspond to processing in RELION[64] and cryoSPARC[65], respectively.

Extended Data Fig 6 Cryo-EM model building for BlaR1.

a, Model of the BlaR1 TM region built using the BlaR(WT) TM map at 3.8 Å resolution. **b**, Side view as in **a**. **c**, Boxed region in **a** showing two phosphatidylglycerol (16:0/16:0) lipids and three phosphates (presumably corresponding to lipid headgroups) modelled in the central chamber between TM regions. Positively charged residues (arginine, lysine, histidine) are shown as sticks. **d**, Model of BlaR1 including sensor domain built using the BlaR1(WT) C2 reconstruction at 4.2 Å resolution. Views in **e** and **f** indicated. **e**, View indicated in **d** showing domain swap mediated dimerization. The domain swap occurs between helices α8 and α9 (*). An AlphaFold2 monomeric model is shown as blue cylinders. **f**, Sensor domain view as indicated in **d** showing fit of X-ray structure (1XA7) in density. **g**, Five AlphaFold2 BlaR1 monomeric models docked in the BlaR1(WT) C2 reconstruction based on the TM region. The different EL1 loop predictions (residues 57-85; low confidence) are indicated with *. **h**, AlphaFold2 models fit in map based on sensor domain. The C-terminal residues 86-96 of EL1 are consistently predicted bound in the sensor domain β-lactam binding groove with increased confidence (pLDDT of 60-75; >70 indicates a confident prediction). **i**, Benzylpenicillin bound to sensor domain (1XA7) overlaid with representative AlphaFold2 model.

Extended Data Fig. 7 Conformational changes associated with autocleavage and β-lactam binding.

a–**c**, Sensor domains modelled into BlaR1(WT) C2 (**a**), BlaR1(F284A) (**b**) and BlaR1(F284A)-ampicillin (**c**) reconstructions. Top front view, bottom side view (indicated with arrow in top panel). Putative signal peptidase cleavage site between Gly330 and Gln331 indicated (**). **d**, Helix α3 position in BlaR1(WT) (green) and BlaR1(F284A)-ampicillin (cyan) representing the two conformational extremes. For comparison of conformational changes, models were first superposed based on residues 200-250 (HEXXH…E motif including re-entrant helix) and reconstructions superposed on these models. **e**, Active site density showing approximate distance between Tyr261 and His201/Asp246. Tyr261 moves closer in the post-cleavage state of the BlaR1(WT) TM reconstruction and in BlaR1(F284A) where the phenylalanine to alanine mutation results in an

empty S1′ pocket. Cryo-EM map shown as mesh contoured at levels to best display protein (grey) and catalytic zinc position (green).

Extended Data Fig. 8 Growth curves for MRSA USA300 in the absence and presence of 2 μg ml^{-1} nafcillin and BlaR1 and BlaI cleavage.

a, The F284A mutation in BlaR1 resulted in an attenuated bacterial survival in presence of nafcillin. Growth assay of SF8300ermS containing pTXΔ with variants of *blaR1* and *blaI* in absence of antibiotics showed similar growth rates among the selected strains. Growth assay of SF8300ermS containing pTXΔ with variants of *blaR1* and *blaI* in presence of 2 μg ml^{-1} nafcillin showed that the F284A functional mutation in *blaR1* (red diamonds) caused an obvious growth defect when compared to the wild type (yellow squares). For each condition three technical replicates and three biological replicates were performed. Data from each biological replicate is graphed separately on each row with growth under 0 μg ml^{-1} nafcillin shown in the left column and growth under 2 μg ml^{-1} shown in the right column. The mean and standard deviation of the technical replicates (n = 3) are shown for each treatment. **b**, Sequence conservation plot generated with the WebLogo server (https://weblogo.berkeley.edu/) of residues flanking the BlaR1 autocleavage site between residues Ser283 (P1) and Phe284 (P1′). **c**, Sequence conservation plot of residues flanking the BlaI cleavage site between residues Asn101 (P1) and Phe102 (P1′). **d**, BlaR1(WT), **e**, BlaR1(F284A) and **f**, BlaR1(USA300) mediate cleavage of BlaI as shown by Western blot. A 2× working concentration of cOmplete protease inhibitor cocktail (Roche) did not prevent BlaI cleavage. However, cleavage was abrogated with the addition of 100 mM EDTA. For panels c, d and e full-length (FL_{BlaI}) and cleaved (CL_{BlaI}) T7-tagged BlaI bands are indicated. BlaI cleavage experiments in **d** – **f** were carried out a minimum of three times with similar results, representative blots shown. **g**, *S. aureus* signal peptidase SpsB mediates cleavage of the BlaR1(WT) sensor domain as demonstrated by SDS-PAGE stained with Coomassie and BOCILLIN FL fluorescently labelled BlaR1 sensor domain. BlaR1 full-length (FL), the BlaR1 C-terminal autocleavage product including residues 284-585 (C-CL), the BlaR1 N-terminal autocleavage product including residues 1-283 (N-CL)

and BlaR1WT sensor domain including residues 332-585 (SD_{BlaR1}) bands are indicated. The BlaR1 G331K (BlaR1(G331K)) mutation in the P1 position of the SpsB cleavage site protects against SpsB mediated cleavage. We predict SpsB mediated cleavage is important for BlaR1 turnover in the absence of β-lactam antibiotic. The molecular weights of selected bands of the protein ladder are indicated in kDa for panels C-F. SpsB cleavage was carried out two times with similar results. For gel source data, see Supplementary Fig. 1.

Extended Data Fig. 9 Mapping of BlaR1 sequence conservation.

Chain A of the BlaR1 dimer is coloured by sequence conservation[81] while chain B is coloured grey. Areas where sequence conservation is uncertain are coloured light yellow. **a**, Overall sequence conservation of the modelled BlaR1 residues. Focus in **b**, the BlaR1 zinc metalloprotease active site and in **c**, the BlaR1 domain swap interface with selected conserved residues shown as sticks.

Extended Data Table 1 *S. aureus* BlaR1 data collection and processing statistics

Supplementary information

Supplementary Information

Supplementary Fig. 1 and Supplementary Tables 1–5.

Reporting Summary

Supplementary Video 1

Overview of BlaR1 dimer structure. One protomer is shown in green ribbon and the other in surface coloured according to electrostatics (Fig. 1d). Residues important to signalling are shown as ball and stick. Positively charged residues close to the membrane are shown as sticks. 00:00–00:14, 360 degree rotation. 00:18–00:34, close up of the active site. The

autocleavage loop is coloured in darker green. 00:40–01:00, sensor-binding site showing EL1 bound in the sensor domain β-lactam-binding groove.

Supplementary Video 2

Structural variation in the BlaR1(WT) sensor domain and zinc metalloprotease active site. Morph between different sensor domain states observed for BlaR(WT). The low-pass filtered BlaR1(WT) C_1reconstruction is shown in grey (solid and transparent), with the model in ribbon with one protomer coloured green and the other coloured grey. 00:11–00:25, close up morph between the two zinc metalloprotease active site states observed in the 3.8 Å BlaR1(WT) reconstruction representing pre-cleavage, with near continuous density along autocleavage loop, and post-cleavage, with clear interrupted density. The distance between Tyr261 and His231 is indicated by a blue dashed line.

Supplementary Video 3

β-lactam-induced conformational changes in BlaR1. Front view morph between BlaR1(F284A) structures without (orange) and with (cyan) ampicillin bound in the sensor domain. Maps are coloured according to Fig. 3 and are shown as solid and transparent surfaces with the structural model shown in ribbon and the protomers coloured green and grey. 00:11–00:15, side view. 00:18–00:22, slabbed (mid-point) side view. 00:25–00:31, animation of β-lactam binding and induced sensor domain movement.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons

licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

About this article

Cite this article

Alexander, J.A.N., Worrall, L.J., Hu, J. *et al.* Structural basis of broad-spectrum β-lactam resistance in *Staphylococcus aureus*. *Nature* **613**, 375–382 (2023). https://doi.org/10.1038/s41586-022-05583-3

- Received: 06 June 2022
- Accepted: 18 November 2022
- Published: 04 January 2023
- Issue Date: 12 January 2023
- DOI: https://doi.org/10.1038/s41586-022-05583-3

Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

Bacterial receptor crucial for antibiotic resistance imaged in detail

Nature Research Briefing 04 Jan 2023

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-022-05583-3

| Section menu | Main menu |